Applications of pattern recognition in medicine

Thesis

APPLICATIONS OF PATTERN RECOGNITION IN MEDICINE

Submitted by

John Richard Thompson

to the Mathematics Faculty of the Open University for the degree of Doctor of Philosophy

April 1985

Date of submission: April 1985

Date of award: 14 February 1986

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

ABSTRACT

This dissertation considers three aspects of the problems involved in applying pattern recognition methods to large medical data sets. The specific examples considered are the diagnosis of depression from subjectively assessed symptoms, the diagnosis of heart disease from a ballistocardiogram and the diagnosis of liver disease from a liver scan.

A new method of variable selection is proposed that involves the elimination of already classified cases. This effectively reduces a multidimensional problem to a series of simpler univariate stages. An alternative to the transformation methods of feature selection is also investigated. This uses the technique known as principal coordinate analysis which, it is argued, should give better results than many of the commonly used techniques.

A method of partial discrimination using tolerance intervals and convex hulls is developed. This work includes a new algorithm, the first to allow the convex hull to be found in any number of dimensions.

Initial misclassification is shown to be a particular problem with the liver scan data and an iterative method for overcoming such errors is suggested. This method is used to modify the linear discriminant function and the nearest neighbour analysis and its performance is studied.

CONTENTS

LIST OF FIGURES

CHAPTER 5:

CHAPTER 6:

CHAPTER 7:

CHAPTER 8:

CHAPTER 9:

CHAPTER 10:

CHAPTER 11:

CHAPTER 12:

CHAPTER 13:

CHAPTER 14:

LIST OF TABLES

1. INTRODUCTION

The main objective of pattern recognition is to bring order to a collection of items by assigning them to classes. These classes are sometimes predetermined, having been defined subjectively, and sometimes the analysis needs to generate its own classes. As an illustration of every day pattern recognition consider the interpretation of speech. We are all accustomed to taking sounds and grouping them into classes where each class represents a different word. Pattern recognition investigates such a process with the long term aim of finding an automatic method of classification.

The development of an automatic pattern classifier is a multi-stage process. It involves the collection of information likely to influence the assignment of items to a class, the sifting of such data to find the most important features, the design of a classifier and the investigation of the performance of the classifier. Throughout, this process is subject to error. Not only is any practical analysis based on only a subset of the features that might have been used but the items from any single class may vary quite considerably. This natural variation within classes is well illustrated by our example of speech recognition. The brain is able to cope with different accents and yet

still recognise a word; the automatic procedures need to be able to do the same.

As the speech recognition example suggests, classification is one of the fundamental mental processes with a long history going back much further than the numerically based automatic procedures developed during the current century. Indeed most sciences pass through an early stage in their development, during which the phenomena under study are grouped and classified.

Bearing in mind the fundamental nature of the process it is not surprising that pattern recognition has found an enormous number of very varied applications. As well as the medical problems that have motivated this work, pattern recognition has been used in areas including biology, psychology, archeology, anthropology and forensic science. Engineers have also been studying pattern recognition for some time; although their language and approach often differ from those of statistics. In their applications they have tended to concentrate on particular types of problem such as speech and character recognition and the interpretation of images, especially those produced by aerial photography.

Medicine offers considerable scope for a statistical approach to pattern recognition. The diagnosis of an illness involves the doctor in selecting and measuring the important symptoms of the patient and the matching of the symptoms with those of the other patients that the doctor has seen or read about. Each disease can thus be seen as

being defined by the set of suffers and characterised by their symptoms. The problem is, of course, complicated by the fact that no two patients will exhibit exactly the same set of symptoms.

A statistical approach to medical diagnosis not only offers the possibility of highly efficient selection and matching of symptoms but also the chance to indicate the probability that the diagnosis is correct. When a computer is available it is also possible to extend the data base of diseases and symptoms to a size far beyond that which any doctor could acquire, even in a life time of practice. Such computerised methods, or 'expert systems' as they have become known, are now the subject of much research and many experiments are already underway to test their performance in the field.

The work in this dissertation has been motivated by the study of three medical data sets. One, from psychiatry, is concerned with the diagnosis of depression from observed behaviour. Another, from cardiology, deals with the diagnosis of heart disease from ballistocardiograms which are waves describing the movements of the body induced by the action of the heart. Finally the third data set, from nuclear medicine, consists of the liver scans of patients with a variety of liver diseases.

These data sets, as well as being of interest in their own right, also have characteristics common to many medical

problems and as such enable us to test general methods of analysis.

The most obvious characteristic of realistic medical classification problems is the enormous number of possible features that one might consider. It is frequently the case with such problems that the number of potential features far exceeds the number of classified cases in the data base. For this reason it is not possible to use all of the features as there would not be enough cases to estimate all of the properties of the features.

Secondly symptoms tend to be highly correlated and of complex distribution. Consequently a sophisticated modelling approach to the data is very difficult. There are comparatively few usable multivariate statistical models and it is necessary to search for non-parametric methods or to argue by anology with what would have happened had the data been multivariate normal.

The third general characteristic of medical data, that is clearly exhibited by our examples, is that of poor initial classification. In setting up relevant data bases it is necessary to rely on the diagnoses of medical specialists and such diagnoses are often far from perfect. There is a danger of ending up by divising a method not for correct diagnosis but for reproducing the mistakes of the specialist.

Our three data sets exhibit all three of these general characteristics. They all contain too many features for

practical use; the largest, namely the liver scans, offering 16,384 values on each of 291 cases. That same data set also best illustrates the third problem, as it was found that only a fraction of the cases had definite usable diagnoses. Finally all three data sets possess the complex, highly correlated, structure that is so difficult to model parametrically.

Because of the large variety that is possible in the structure of medical data sets, it is very difficult to compare methods of analysis, for it is not possible to be sure whether success is due to the method or to some inherent characteristic of that particular data set. The need for a standard data set on which competing methods might be tested leads to the compromise of using multivariate normal data. Clearly it is not the intention to devise a method that will out perform the well known optimal procedures, but it might be argued that unless a method can cope with such well behaved data it is unlikely to be generally applicable. Indeed it would be desirable for the method under test to reduce to the optimal procedure in the case of normal data.

This dissertation starts with a general review of the present state of statistical pattern recognition and then goes on to introduce the three data sets. Conventionally pattern recognition is broken down into two stages, namely the selection of the important features and the design of

the classifier. However with complex data sets such as those under consideration an extra stage is neccessary because of the many ways of looking at the original problem. Thus, for example, the ballistocardiograms could be viewed as sets of 100 amplitudes, or just as reasonably as sets of peaks and troughs. It is only after this initial decision has been made that the features are defined and the important ones can then be chosen. For this reason we consider, in chapter four, the various ways in which the data might be characterised. The characterisations include a new and successful way of modelling the ballistocardiograms as superimposed damped harmonic waves.

A detailed review of feature selection, in chapter five, is followed by the consideration of two new approaches to the problem. The first new approach seeks to reduce the complicated multivariate selection problem to a series of more managable univariate problems by looking at the distributions conditional upon unhelpful values for those variables already selected. Thus, the best variable is selected first and the next variable is then chosen on the assumption that the observed value of the first will turn out to be unhelpful.

The second approach uses the standard multivariate technique of principal co-cordinates. This technique is well known outside the field of feature selection and has many properties that make its use in feature selection desirable. These properties include the fact that the well

established techniques of principal components and canonical variate analysis may be viewed as special cases of principal co-ordinates. It will be shown how this technique may be modified to offer many other powerful possibilities all of which are tested on the medical data sets.

Methods of defining linear classifiers are considered in chapter eight and in chapter nine we review the method of partial discrimination in which, as well as the possibility of classifying the case, one adds the option of leaving it unclassified.

Chapter ten introduces a new method of non-parametric partial discrimination which is applied to two of the data sets. The method is then extended into higher dimensions. The extension is based on the convex hulls of the samples and in order to facilitate this a new algorithm is proposed for finding the convex hull of a p-dimensional sample. This proved necessary as those algorithms currently available are restricted to at most three dimensions.

As was mentioned earlier the liver scans are plagued by poor initial classification. When the diagnoses originally supplied were checked against the information that subsequently became available it was found that a surprising number of the original cases appeared to have been misclassified.

Methods for coping with initial misclassification are reviewed in chapter thirteen and a new procedure is suggested

and tested. Finally the liver scans are considered spearately and in detail.

The new methods suggested in this dissertation have varying degrees of success when measured against the three data sets that motivated them but all appear to offer useful additions to the battery of available techniques.

2. A Review of Discrimination, Clustering and Pattern Recognition

2.1 A Brief History

The borders between discrimination, clustering and pattern recognition are not at all clear cut, but the term discrimination is usually reserved for situations in which one has sets of previously classified cases, referred to as training sets, which are used in the design of a scheme for classifying future cases. Clustering on the other hand, refers to the problem of dividing unclassified cases into groups without any previous class information available as a guide.

Pattern recognition is much more difficult to define. It is a term that is rarely, if ever, used by statisticians but for engineers and computer scientists it covers discrimination, clustering and a whole lot more. Indeed it is used as a generic term for any analysis that involves the search for similarities between the characteristics of different cases and the use of those similarities to recognise a case as belonging to a particular class. This

definition, similar in essence to those used by reviewers of the field, such as Fu(1980), Kanal(1974) and Verhagen(1975), is wide enough to stretch from the simplest exploration of a data set using plots to the most sophisticated discriminant analysis and yet also includes a number of non-statistical techniques.

Whilst there is now considerable overlap between the three fields of study they each had there own separate origin. The first, and in many ways most important contribution to discrimination was made by Fisher(1936). His largely intuitive method was based on the idea that two classes are most easily distinguished if one looks in the direction that maximises the ratio of the difference between their means to their common standard deviation. Fisher's result was later obtained using probabilistic and decision theoretic analyses, together with the assumption of multivariate normality, by Welch(1939) and Wald(1944), and this technique remains the most widely used. Non-parametric methods began to be considered in the early 1960's following a paper by Parzen(1962) and distributions other than the normal, including some discrete cases, have now been studied.

The Bayesian school of statisticians were lead by Geisser(1964) in their study of discrimination. The failure of their methods to gain widespread acceptance is partly a result of the added complexity of the analysis and partly

due to the continuing debate over the validity of the general approach. The extra problems of specifying sensible prior distributions and the lack of commonly available software are together suffient to put off all but the specialists in this field.

Clustering can be traced back to the 1930's when Zubin(1938) attempted to apply simple ad hoc rules to specific data sets. Even more than discrimination, the field of clustering is dependent on the use of computers and little real progress was made until the 1960's when computers began to be widely available. At that time there was an explosion of work in the field with countless papers being written each suggesting its own algorithm. The suggestion of ad hoc clustering procedures continues to this day with very little in the way of underlying theory, and indeed, very few guidelines to help a user choose between the multitude of methods.

Being more vaguely defined it is harder to say just when pattern recognition was first studied but the term seems to have been coined by Sammon(1968) who wrote on "On-line Pattern Analysis and Recognition Systems". The literature soon sucked in the statistical work adding to the theory and introducing the methods to many new and diverse applications. Because pattern recognition was being studied by people from very different backgrounds, such as computer science and engineering, this naturally led to a whole new

set of approaches.

The major alternative to the statistical approach is the syntactic or structural method, some of the early theory of which predates Sammon's 1968 paper. In the mid 1950's Chomsky and others started to develop mathematical models for the grammars of natural language; work which led to the desire to program computers to understand and to translate. The analogy between these aims and those of pattern recognition was siezed on by many workers. The patterns of interest, that is to say the sentences are divided up, or segmented, into basic components called words or morphs. These basic units would typically be more complex than the single measurements used in a statistical analysis. For example, if one were analysing handwriting the morphs might be the various strokes. The grammar of our problem is then the way in which these basic units are put together to make meaningful patterns and the syntax is the set of rules that dictate which morphs may follow one another.

By these means the patterns or sentences are composed of basic units according to specified rules and we may think of the whole process as being represented by a tree diagram, with the starting unit at the base it fans out according to the possibilities of the syntax to end with the complete pattern. Again taking a term from language theory, a pass through such a tree is known as a parse.

The structural method is still being developed and

extended and recently other essentially non-statistical approaches have been tried. These new methods have brought in ideas from such fields as fuzzy set theory, variable-value logic, category theory and relation theory. Whilst each approach is worthy of study in its own right, we will concentrate on the statistical approach in the rest of this review since that is the concern of the rest of this thesis.

2.2 Literature and Software

The literature on all these areas, but especially on pattern recognition, is enormous. The best known text books dealing primarily with discriminant analysis are those by Goldstein and Dillon(1978), Hand(1981) and Lachenbruch(1975), although it is dealt with from a Bayesian view point in Aitchison and Dunsmore(1975), and from an information view point in Kullback(1959). Clustering is described fully in Anderberg(1973), Everitt(1980), Hartigan(1975) and Jardine and Sibson(1971). There are even more standard texts on pattern recognition, Fu(1980) in a review lists twenty, the most widely quoted being Andrews(1972), Batchelor(1974), Becker(1971), Chen(1973), Chein(1978), Duda and Hart(1973), Fu(1968), Fu(1974), Fukunaga(1972), Mendel and Fu(1970), Patrick(1972), Tou and Gonzalez(1974), Ullman(1973), Van Ryzin(1977) and Wantanabe(1972). Since that review notable publications include Devijver and Kittler(1982) and Hand(1981). There are also numerous collections of papers and books dealing which specific applications such as image analysis.

All three fields have been regularly reviewed in the

literature, the best general reviews being by Lachenbruch and Goldstein(1979) on discriminant analysis; Cormack(1971) and Blashfield and Aldenderfer(1978) on clustering, and Kanal(1974) and Fu(1980) on pattern recognition.

The factor that probably has the greatest influence on the methods used in practice is the availablity of software. Most large statistical packages include programs for discrimination and clustering, most common in Britain being SPSS, GENSTAT and BMD-P. The scope of the analysis offered by these packages is however quite limited, for example each offers only multivariate normal discrimination.

A few specialist packages, such as CLUSTAN are widely available and several text books have included listings of programs for specific analyses. These include Goldstein and Dillon(1978) and Hartigan(1975). Programs are also published regularly in such journals as 'Applied Statistics' and 'The Computer Journal'. However these programs have considerably less influence on the types of analysis routinely performed. It is not at all unusual to find the multivariate normal distribution assumed simply because of the availability of the software.

There is no major package aimed specifically at pattern recognition; this is not surprising when one considers the vastness of the field and the tendency to require methods specific to a particular problem.

2.3 Discrimination

In this section we will review the major aspects of discrimination and statistical pattern recognition based on classified training sets. Any such analysis can be subdivided into three stages, the selection of the variables or features that best discriminate between the classes, the design of the classifier and finally the assessment of the whole scheme. These three parts do, of course, interrelate. For example, the variables chosen will dictate the choice of classifier and the assessment criterion will influence both the choice of features and the design of the classifier.

Feature selection and classifier design are reviewed in some depth in chapters five and eight respectively. Our intention here is merely to outline the basic results so as to place the rest of the thesis into a general context.

Following the notation set out in appendix I, suppose that we observe a vector of measurements $\underline{x}$ from each of m classes, (w_j), $j=1,\ldots m$, and that the densities of such measurements is $f(\underline{x}|w_j)$ with a priori probabilities $P(w_j)$. A classifier will divide the space S into mutually exclusive and exhaustive subsets (S_j), $j=1,\ldots m$, and the associated classification rule will assign $\underline{x}$ to class w_j if $\underline{x} \in S_j$.

The most commonly used criterion for designing a classification scheme is the minimisation of the total cost of classification given by,

$$\sum_{j=1}^{m} \sum_{i=1}^{m} \int_{S_i} c_{ij} \, f(\underline{x} \mid w_j) \, P(w_j) \, d\underline{x}$$

where c_{ij} is the cost of classifying a case from w_j into w_i. These costs are usually difficult to evaluate and it is common practice to assume that $c_{ij}=1$ for all i and j, and $c_{ii}=0$ for all i, so that the criterion reduces to the minimisation of the probability of error.

This is not the only possible criterion. One might, for example, choose to minimise the maximum probability of misclassification over the m classes.

Minimisation of the total probability of misclassification leads to the so called Bayes classifier, given by,

$$S_j = \{\underline{x} \; ; \; f(\underline{x} \mid w_j) \, P(w_j) > f(\underline{x} \mid w_i) \, P(w_i), \; i \neq j\}$$

The name derives from the work of Bayes and specifically his theorem which states that,

$$P(w_j \mid \underline{x}) = \frac{f(\underline{x} \mid w_j)\, P(w_j)}{\sum_{i=1}^{m} f(\underline{x} \mid w_i)\, P(w_i)}$$

Consequently S_j is the region that maximises $f(\underline{x}|w_j)P(w_j)$ or equivalently which maximises $P(w_j|\underline{x})$. The later being the probability of class w_j given the data $\underline{x}$.

In theory, so long as the probability of misclassification is accepted as the best criterion, then the discrimination problem is solved. However in practice there are many problems linked to the need to estimate $f(\underline{x}|w_j)$ and $P(w_j)$. Unfortunately there are few multivariate families of densities that are both tractable and realistic. The result is that one is forced either to assume multivariate normality or to estimate the density non-parametrically; the latter being a complex problem in itself if the number of dimensions is large. These difficulties have lead to the suggestion of numerous ad hoc procedures which give sensible if not optimal results.

The one density that is tractable is the p-dimensional multivariate normal distribution which has a density,

$$|2\pi \underline{\Sigma}|^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}(\underline{x} - \mu)' \Sigma^{-1} (\underline{x} - \underline{\mu})\}$$

If the classes share a common covariance matrix $\underline{\Sigma}$ then the Bayes classifier will divide the space S into subsets S_i using p-dimensional planes, and when there are two classes an equivalent classification can be obtained using the rule based on,

$$|\underline{x} - \tfrac{1}{2}(\underline{\mu}_1 + \underline{\mu}_2)|' \underline{\Sigma}^{-1} (\underline{\mu}_1 - \underline{\mu}_2) <> |\ln|P(w_2)/P(w_1)|$$

If the maximum likelihood estimates are substituted for $\underline{\mu}_i$ and $\underline{\Sigma}$ then one obtains the result obtained by Fisher(1936) using heuristic methods.

Without the assumption of equal covariance matrices the classifier is quadratic rather than linear. The boundaries being p-dimensional hyperquadrics. Generally linear classifiers are simpler to estimate and easier to interpret so that much interest has been shown in finding linear discriminant functions for situations in which they are not actually optimal.

Anderson(1972) looked at the problem in a slightly different way by attempting to model $p(w_j \mid \underline{x})$ using a logistic function. Thus one takes,

$$P(w_j|\underline{x}) = \exp(\underline{a}_j' \underline{x}) P(w_m|\underline{x}) \quad j = 1, \ldots m$$

where

$$P(w_m|x) = \frac{1}{1 + \sum_{i=1}^{m} \exp(\underline{a}_i' \underline{x})}$$

This form can be shown to be exact in several situations including, (i) multivariate normal distributions with equal covariance structures,

(ii) multivariate independent 0,1 variables,

(iii) multivariate 0,1 variables, following a log linear model with second and higher order effects equal to zero

(iv) any combination of (i) and (iii).

The parameters $\underline{a}_j$ can be estimated by the method of maximum likelihood although such estimates can suffer from considerable bias. Anderson and Richardson(1979) have extended this work suggesting ways of reducing such bias.

So far we have looked at the discrimination problem from a classical view point but there is one important alternative school of statisticians, the Bayesians, who would consider the problem somewhat differently. A Bayesian analysis is distinguished by a subjective view of probability and the assumption of prior distributions to model the statisticians beliefs about the parameters before the data are collected.

In the case of discriminant analysis the Bayesian would first require a density $f(\underline{x}|\underline{\theta})$ for the data and a prior distribution, $r(\underline{\theta})$ for the parameters. Given the sample of

data $\underline{x}_i$, i=1,...,n he would then wish to modify his views about the parameters and this would be done according to Bayes theorem. That is the posterior distribution would be given by,

$$r(\underline{\theta}|\underline{x}_i,\ i=1,\ldots,n) = \frac{L(\underline{\theta}|\underline{x}_i,\ i=1,\ldots,n)\,r(\underline{\theta})}{\int L(\underline{\theta}|\underline{x}_i,\ i=1,\ldots,n)\ r(\underline{\theta})\,d\underline{\theta}}$$

where L denotes the likelihood.

This information can then be used to set up a predictive distribution by which future values $\underline{x}_u$ can be classified. Thus the predictive distribution is given by,

$$f(\underline{x}|\underline{x}_i,\ i=1,\ldots,n) = \int f(\underline{x}\ |\underline{\theta})\ r(\underline{\theta}|\underline{x}_i,\ i=1,\ldots,n)\ d\underline{\theta}$$

This approach was first introduced by Geisser(1964) and is described in some detail in the book by Aitchison and Dunsmore(1975).

Bayesian analyses are frequently criticised for their subjectivity. The main problem in practice is that it is often difficult to specify a sensible prior, especially when it is realised that no prior can be objectively tested. Certainly the Bayesian analysis tends to be more complex and

usually requires that the priors selected take simple forms before it is possible to give expressions for the predictive distributions.

Aitchison, Habbema and Kay(1977) published a comparison between the predictive, or Bayesian approach and the estimative or classical. On both theoretical and practical grounds, supported by a small simulation, they came down firmly in favour of the predictive approach. However, Moran and Murphy(1979) in an article reconsidering the problem, showed that the advantage of the Bayesian approach could be removed by the use of an unbiased estimator of the linear discriminant function. That is by using,

$$\frac{n_1 + n_2 - p - 3}{n_1 + n_2 - 2} (\underline{\bar{x}}_1 - \underline{\bar{x}}_2)' \underline{S}^{-1} | \underline{x} - \tfrac{1}{2}(\underline{\bar{x}}_1 + \underline{\bar{x}}_2) | + \tfrac{1}{2}p(\frac{1}{n_1} - \frac{1}{n_2})$$

They also note, as do Aitchison and Dunsmore that the Bayesian predictive result can itself be derived by classical likelihood ratio methods as described in Anderson(1966).

The choice between classical and Bayesian statistics is not an easy one and whilst it is probably true that the Bayesian approach to statistics in general will become the most widespread in use, there remain many problems in its application especially in the specification and validation of priors. As a result it is probably too early to take on

the Bayesian methods.

Returning then to the classical view of the problem, another way of obtaining Bayes classification for multivariate normal data when the covariance matrices are equal and the a priori probabilities are equal, is to use the Mahalanobis distance from $\underline{x}$ to the class means $\underline{\mu}_i$, that is,

$$(\underline{x} - \underline{\mu}_i)' \underline{\Sigma}^{-1} (\underline{x} - \underline{\mu}_i)$$

From the form of the multivariate normal density it will be seen that Bayes classification is obtained if one assigns $\underline{x}$ to the class to which it is closest. Once again the simplicity of the result has lead to the suggestion of many other distance measures that are either robust or applicable to other distributions. It is usually the case that such distance measures give sensible if not optimal results.

The only non-normal distribution that has been studied extensively is the multinomial. This most general model for discrete data assumes that there are k cells each representing a value of the variable. $P(K_i|w_j)$ is then the probability of a case from class w_j falling in cell K_i. The Bayes classification procedure in this case will depend upon the relative sizes of $P(w_j)P(K_i|w_j)$ for different classes w_j. These cell probabilities could be estimated in the most general case by the proportion of each class that fall in that cell.

This is a potentially important approach because any set of variables could be discretised, a procedure that was studied by Cochran and Hopkins(1961). However the problem is once again one of estimation, for unless the training sets are very large in comparison to the number of cells, we will be unable to obtain stable estimates.

Hills(1971) suggested that the variances of the simple proportion based estimators could be reduced by using information from neighbouring cells. Thus if n_j is the number of cases in the training set that come from class w_j, and n_{ij} of them are found in cell K_j, then the maximum likelihood estimate of the cell probabiliy is,

$$P(k_i|w_j) = \frac{n_{ij}}{n_j}$$

If, however, a set A of neighbouring cells is defined then one might instead take,

$$P(k_i|w_j) = \frac{n_{ij} + \sum_{l \in A} n_{lj}}{n_j}$$

giving a smaller variance at the expense of some bias.

Various models have been proposed for the structure of multinomial data sets; mostly these enable one to make simplifying assumptions, such as no interaction between the variables, thus reducing the number of parameters that have to be estimated. Models of this type include those of Bahadur(1961), a log-linear model proposed by Berkson(1955), a logistic model proposed by Day and Kerridge(1967) and Cox(1970) which, as we have seen, was later incorporated into the work of Anderson(1972), and a number of models based on orthogonal polynomials. A good general survey of discrete methods is found in Goldstein and Dillon(1978)

The lack of other parametric alternatives to the multivariate normal has lead to research along two lines. Into the robustness of the normal based results and into non-parametric methods.

Robustness is considered later in chapter eight. We will see that such studies concentrate on either, the effect on the results derived from normal theory when the data follow some non-normal distribution, or the effect when normal data are contaminated by values from a second density, or the effect of assuming equal covariance structure when the assumption is not justified. One might, however, include under this heading the effect of using training sets that contain misclassified cases. This topic is covered in chapter thirteen.

Clearly near Bayes classification would be possible if we could obtain a good non-parametric estimate of the densities. The literature on such estimates is now extensive, see Wegman(1972) for a review. The two main approaches are based on Kernel estimates and orthogonal series.

Kernel estimates for univariate densities take the form,

$$f(\underline{x}) = \frac{1}{n\,h(n)} \sum_{i=1}^{n} k\left(\frac{x_i - x}{h(n)}\right)$$

where $x_1,\ldots,x_n$ are the n values in the training set, K is a function satisfying regularity conditions given by Parzen(1962), and h(n) is a function that controls the smoothness of the estimate. The main problem that remains unsolved with this method is the reliability of the estimate. Although the choice of the function K does not seem critical, the value of h(n) can alter the appearance of the estimate quite dramatically. Further, multivariate generalisations require enormous amounts of data if they are to give reliable results.

The corresponding form for estimates based on orthogonal series is,

$$f(x) = \sum_{j=0}^{q(n)} a_j \, g_j(x)$$

where $g_j(x)$ is the series of normalised Hermite functions and

$$a_j = \sum_{i=1}^{n} g_j(x_i)/n$$

Once again the problem is to choose q(n), the factor that controls the smoothness of the estimate and to find ways of extending the method to multivariate data without the requirement of huge data sets.

One other non-parametric density estimate is in common use and that is based on the idea of taking a volume V around the unclassified point $\underline{x}$. The size of the volume is chosen so that V contains k of the points from the training sets. If then k_j of those points come from class w_j then we might use the estimate,

$$f(\underline{x}, w_j) = \frac{k_j/n}{v}$$

so that,

$$P(w_j|\underline{x}) = \frac{f(\underline{x},w_j)}{\sum_i f(\underline{x},w_i)} = \frac{k_j}{k}$$

This idea leads naturally to the k^{th} nearest neighbour rule that has received so much attention since its original suggestion by Fix and Hoges(1951). According to this method the unclassified case $\underline{x}$ is placed in the class most common amongst the K nearest neighbours.

With any of the methods of classification that lead to density estimates we might in turn estimate the probability of misclassification or error of the scheme by,

$$1 - \sum_{j=1}^{m} \int_{S_j} f(\underline{x}|w_j)\, P(w_j)\, d\underline{x}$$

Unfortunately this estimate is bound to give a falsely optimistic view because the boundaries of S_j will have been drawn to accommodate the estimated densities $\hat{f}(\underline{x}|w_j)$ and not the actual densities $f(\underline{x}|w_j)$.

Similarly if one applies the classification scheme to the

training data used to create it, one is bound to get an optimistic view of the scheme's performance; for the scheme is geared to the sample and not to the population.

This apparent error rate was first suggested as a way of assessing classifiers by Smith(1947), although it is now only used with great caution. Lachenbruch(1965) suggested the use of an alternative 'leave-one-out' method. According to this approach each case is left out in turn and the boundaries of the classification scheme are estimated from the others. The omitted case is then used to test the scheme. By repeating this process with each case one gets an estimate of the error rate that is only slightly pessimistically biased.

Of the many simulated studies of the error rate the results of Fukunaga and Kessel(1971) are both typical and illustrative. They took two eight dimensional normal distributions with unequal covariance matrices and an actual probability of error of 0.019. The results of the leave-one-out method of error rate estimation were then compared with the estimates based on estimated hyperquadrics. Table 2.1 shows the results.

Table 2.1

Apparent and 'leave-one-out' estimates of the error rate for various sample sizes.

Sample sizes	Error Rates	
equal	Apparent	Leave-one-out
100	.0144 (-24%)	.0215 (13%)
200	.0156 (-18%)	.0200 (5%)
400	.0183 (- 4%)	.0197 (4%)

percentage bias shown in brackets

Whilst the apparent error rate is seen to be falsely optimistic, the leave-one-out method is falsely pessimistic, although not to such a great degree. In both cases the bias reduces as the sample size increases.

Many other methods have been suggested for reducing the bias in the special case of multivariate normal distibutions with equal covariance matrices; Lachenbruch and Mickey(1968) giving a comparison. Unfortunately such approaches are not easily generalised for use with non-normal data.

Given the classifier the problem of feature selection reduces to one of looking at the probability of error associated with each subset of the available variables. However, as we have noted, it is not easy to estimate the probability of error, and if the number of potential variables is large then the number of possible subsets to be considered may be enormous. To give some idea of the scale

of the problem with forty possible variables to choose from there are some million million possible subsets to be considered.

Methods for searching out good subsets without exhaustive search are detailed in chapter five, together with a review of the various ad hoc techniques in common use for defining a good subset or function of the variables.

The question of how many feature to include is further complicated by the finite sizes of the training sets. It is true that as one increases the number of features so the apparent error rate must decrease but there is likely to come a point when one has so many features that the samples are no longer large enough to given reliable estimates of the multidimensional densities. Consequently the classifier will be poorly estimated and the actual error rate will start to increase.

This apparent paradox was noted quite early in the history of pattern recognition and has been studied using simulation methods by a number of workers. The most notable studies being those by Liddell(1977), Van Ness(1979), Van Ness and Simpson(1976), Jain and Walker(1978) and El-Sheikl and Wacker(1980). Unfortunately there is no simple rule of thumb to tell one how large the training sets need to be in order to justify a particular number of feature. The solution to that problem being dependent on the structure of the data.

This whole problem is further complicated by the fact that the same phenomenon can be observed as a result of the use of an inappropriate model for the density. The reason being that the false model may be a better approximation in some dimensions than others. Hand(1981) quotes an example of just such a case in which a one dimensional classifier out performs a two dimension classifier. Thus the problem is not restricted to large scale studies.

2.4 Clustering

Clustering, or unsupervised analysis, involves the derivation of a classification rule from a training set that has not itself been classified. The analysis must therefore, be capable of defining the groups or clusters for itself.

Despite the large amount of published work in this field cluster analysis remains a collection of largely unrelated procedures. Indeed the fact that the large majority of papers on this topic deal with applications emphasises the importance of the methods to practitionersand the lack of an underlying theory.

The methods in common use include two main types, namely hierarchical and iterative partitioning. The former uses a measure of distance to build up clusters of close elements. At each stage individual cases or already existing clusters are merged if they are the closest. In this way a hierarchy of association is produced. Clearly the definition of the distance between two groups is a matter for debate and many definitions have been suggested. For instance, one might use the distance between the closest cases (nearest neighbour), or the distance between the furthest separated cases (furthest neighbour), or any one of a number of average

distances. It is this arbitrariness in the choice of the distance that is to be used to measure the separation of cases and groups that is the root of much of the dissatisfaction with the overall approach.

Iterative partitioning requires that one specifies the number of groups that one expects and an initial partition of the cases. Cases are then moved between the groups with the object of optimising some measure of the degree of clustering. For example, one might try to maximise the average Mahalanobis distance between groups. A full account of these and other methods in common use may be found in any of the standard texts referred to in section 2.2.

The lack of a theoretical framework for cluster analysis means that there is little hope of being able to give guide lines for deciding which of the many clustering methods should be used on any particular problem. As things stand at the present the best advice seems to be that one should try several methods and only believe in those clusters that appear in all of the analyses.

One approach related to clustering that does have some theoretical basis is the method of mixtures. According to this procedure a model is proposed for the classes suspected of being represented within the data. The unclassified data are then treated as a mixture from those distributions and hence it is possible to specify the likelihood of the sample. The parameters of the mixture can then be estimated

by maximising the likelihood. In chapter twelve this approach is applied to data suspected of containing initially misclassified cases.

More detailed reviews of the aspects of pattern recognition relevant to the work of this thesis will be found in subsequent chapters, but next we move on to consider the data sets that motivated the proposed methods.

3. THE DATA SETS

3.1 Introduction

Three sets of medical data will be analysed in this thesis. As well as being of interest in their own right, they will also serve to illustrate the main problems of pattern recognition and are representative of three major areas of application. One set, the psychiatric data, consists of symptom ratings; another set, the ballistocardiograms, consists of waveforms describing movements of the body induced by the beating of the heart; and the final set, the liver scans, consists of radionuclide images.

The psychiatric data set contains subjectively ordered categorical ratings for symptoms measured on patients suffering from one of five types of depression. The ballistocardiograms record the movements of the body resulting from the beating of the heart and the consequent flow of blood and were obtained from samples of healthy and

unhealthy subjects. The largest data set is that containing the liver scans; these were recorded as part of the routine work of a hospital and consist of the radionuclide images of patients with a variety of actual or suspected diseases.

All three data sets present problems of feature selection and classifier design, but the liver scans pose the added problem that they are difficult to interpret, even for experienced clinicians, and consequently the diagnoses originally given are somewhat doubtful.

As one might expect with such data sets, the normal based theory that dominates in pattern recognition is not strictly applicable. However it is often of interest to see how a proposed method of analysis would perform on normal data, for this is often the only way that different methods can be compared. To this end several sets of multivariate normal data were generated and used for testing the proposed methods.

3.2 Data set 1: The Psychiatric Data

This data set was collected by Caetano(1980), a psychiatrist interested in the differentiation of five types of depressive illness. The sample of acute patients was taken from four catchment area, Cambridge, Peterborough, Northampton and Stevenage. Patients were selected on the basis of a medical or nursing report that there had been a mood change towards depression and or anxiety, but patients who, at interview, showed signs of organic impairment, schizophrenia, manic psychosis or any non-affective neurotic syndrome were excluded.

The total sample size was 152, and included patients ranging from 20 to 80 years of age, with roughly twice as many women as men.

All patients were interviewed for about 50 minutes within their first week of admission to hospital, that is to say, before their treatment began. The Present State Examination (PSE) was used as the basis for the scoring of the patients. PSE is a well established and reliable means of scoring affective disorders and the manual was followed in its entirety. According to this scheme each symptom is scored,

0 not present in previous month

1 present in a moderate form during the last month

2 present in a severe form during the last month

8 unsure

9 not asked or not applicable

To obtain a natural progression of scores the psychiatrist decided to rescore '8' as '0.5'. Thus each symptom was scored 0, 0.5, 1., 2 or 9.

The psychiatrist then reduced the number of symptoms by,

(i) excluding those symptoms present in less than 10% of the sample.

(ii) excluding those symptoms with a high proportion of missing values.

(iii) compressing some related symptoms into single items. Where symptoms were combined a further point '3' was added to the scoring scale. The complete list of symptoms used is shown in table 3.1.

The 152 patients were also classified into one of five classes of affective disorder, namely,

1. Psychotic depression (n=13)
2. Retarded/Agitated depression (n=32)
3. Endogenous depression (n=19)
4. Neurotic depression (n=39)
5. Anxiety state (n=41)

This left four cases about which the psychiatrist was unsure, although he did comment that this group probability consisted mostly of 'anxiety state' patients.

1 worrying
2 tension pain
3 tireness
4 muscular tension
5 restlessness
6 hypochondriasis
7 nervous tension
8 free-floating autonomic anxiety
9 panic attacks
10 situational autonomic anxiety
11 autonomic anxiety on meeting people
12 specific phobias
13 anxiety avoidance
14 inefficient thinking
15 poor concentration
16 brooding
17 loss of interest
18 depressed mood
19 hopelessness
20 suicidal plans or acts
21 morning depression
22 social withdrawal
23 self depreciation
24 lack of self confidence
25 ideas of reference
26 guilt ideas of reference
27 pathological guilt
28 loss of weight
29 delayed sleep
30 subjective retardation
31 early waking
32 irritability
33 derealisation
34 depersonalisation
35 subjective loss of effect
36 agitation
37 observed anxiety
38 observed depression
39 delusion or hallucination (*)
40 hysterical symptoms (*)
41 obsessive symptoms (*)

(*) amalgamated symptoms

Table 3.1

List of Symptoms

3.3 Data Set 2: The Ballistocardiograms

Physicians in the nineteenth century were aware that the pumping of the heart and the resulting movement of the blood produces a detectable movement in the body of an otherwise still subject. Gordon(1877) used the much quoted analogy between the movement caused when the blood is forced from the heart and the recoil of a gun. Then, shortly after the turn of the century, Henderson(1905) devised the first instrument, in the form of a swinging bed, for measuring that recoil.

The systematic study of the phenomenon had to wait until the time of the second world war when pioneers such as Isaac Starr were spurred by the discovery that the patterns of movement produced in young healthy subjects showed a regular reproducible waveform and further that this waveform differed greatly from that observed from patients with heart disease.

The trace of the bodies movements, known as a ballistocardiogram or Bcg, was at that time still being measured by instruments consisting of a bed suspended from four parallel cables. Interest was, at that time, generally confined to movement in the head to foot direction.

By the 1960's beds had been developed that floated on air and which measured the rotation of the body as well as movement in three perpendicular directions; see, for example, Cunningham and Smiley(1961). The first instrument to measure the ballistocardiogram of a seated subject was

produced by Harrison and Talbot(1967) and now measuring devices are being linked to microprocessors, so offering the possibility of real time analysis.

Interest in the ballistocardiogram has undergone a marked decline since those early years. There was an initial rush of enthusiasm when it was thought that the Bcg might be used to predict people likely to suffer a heart attack. When this early promise failed to materialise, interest was largely lost and although work has continued, the ballistocardiogram has never been a generally accepted diagnostic tool.

In order to appreciate the significance of the peaks and troughs of the ballistocardiogram it is necessary to have a rudimentary understanding of the workings of the heart. As is well known, the heart is divided into four chambers, a right and left atrium and two corresponding ventricles. Blood returning from being oxygenated in the lungs passes into the left atrium which, after filling, contracts (atriole systole) and forces blood into the left ventricle. This chamber in turn contracts (ventricular systole) so forcing the blood out of the heart and into the main artery which is known as the aorta. From there the blood passes around the body returning to the heart via the right atrium. The smaller right side of the heart pumps in time with the left forcing the blood into the right ventricle and then out to return to the lungs. The time during which the chambers are contracting is known as the systolic phase and the time

during which they relax to refill with blood is known as the diastolic phase. The entire cycle is repeated between 60 and 100 times per minute in a healthy individual.

The ballistocardiogram is by no means the only way of following this cycle. Much more commonly used as diagnostic indicators are the sounds produced by the beating heart and the pattern of electrical activity that fires the heart, known as the electrocardiogram or Ecg.

Figure 3.1 shows a typical healthy head to foot ballistocardiogram, with its main features denoted in the conventional way by the letters G to P. Figure 3.2 shows the corresponding Ecg and typical measurements of the pressures in the four chambers.

The small early deviations in the ballistocardiogram are associated with the atriole systole. The movements of the body during the systolic phase of the ventricles produce the peaks G to K and the later peaks, L onwards, correspond to the diastolic phase.

Figures 3.3 and 3.4 show the ballistocardiograms of two actual subjects, one healthy and the other pathological. It will be seen at once how the unhealthy individual's Bcg differs markedly from the general pattern shown in figure 3.1.

Many attempts have been made to find the features of the ballistocardiogram that carry the important diagnostic information. The two main types of feature that have

received the bulk of the attention being the amplitudes of the peaks and the intervals between them. Representative papers are those of Jaramlokova-Nicolova(1973) and Grebenarov(1973). Jaramlokova-Nicolova found that the amplitudes and timings of the H and J waves were the important features for distinguishing between healthy subjects and those suffering from coronary heart disease, and Grebenarov, when comparing healthy subjects with those suffering from deteriorated myocardial function, found that the H, J and K amplitudes and intervals were important. These papers are typical in concluding that the large peaks of the systolic phase are the best diagnostic indicators.

Our particular data set was collected at the University of Strathclyde, on a highly sensitive instrument comprising of a light carbon fibre bed floated on air. One hundred and thirty-one head to foot ballistocardiograms were measured by simultaneously recording the Bcg and Ecg for periods of 10 seconds. As is the custom in this type of study, the peak of the QRS wave of the electrocardiogram was taken as the starting point for each cycle. The resulting waves are then averages over several beats within the 10 seconds. They have been digitised at 100 equally spaced points and scaled so that the amplitudes lie in the range -100 to 100.

Even after the averaging the waves still showed some signs of noise and so they were pre-proccessed using a Fourier filter. A low pass threshold set at 0.1 Hertz removed the effects of breathing, and a high pass filter set

at 30 Hertz took out the mains noise.

The ballistocardiograms have been divided into two training sets consisting of 81 normal and 50 pathological cases. Unfortunately there is no further information available on the subjects, although information on age, weight and detailed diagnosis would have been helpful. The most significant missing information is however, simply the indentities of the subjects; the data sets actually consist of repeat measurements on a comparatively small number of subjects. It is believed that there were about ten healthy subjects and half a dozen unhealthy ones. This fact has obvious implications for any analysis of the data.

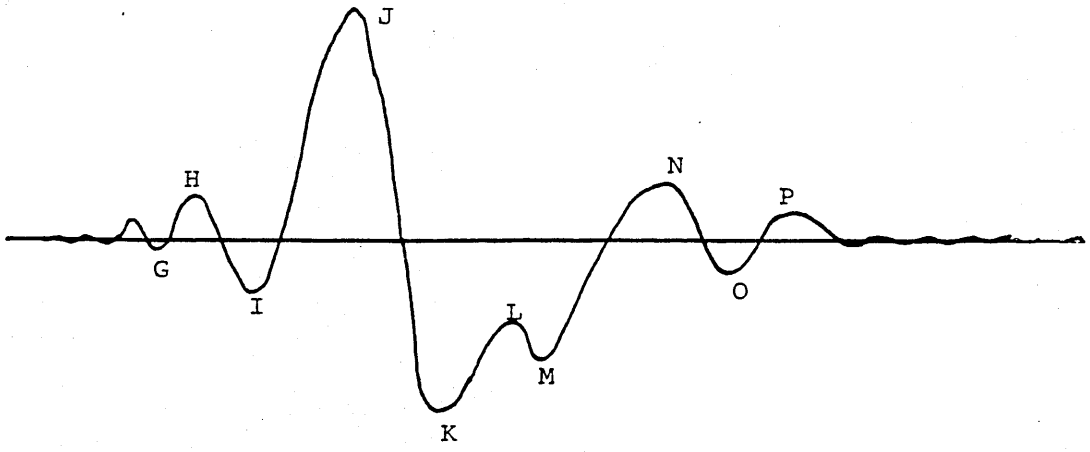


figure 3.1

A typical healthy ballistrocardiogram

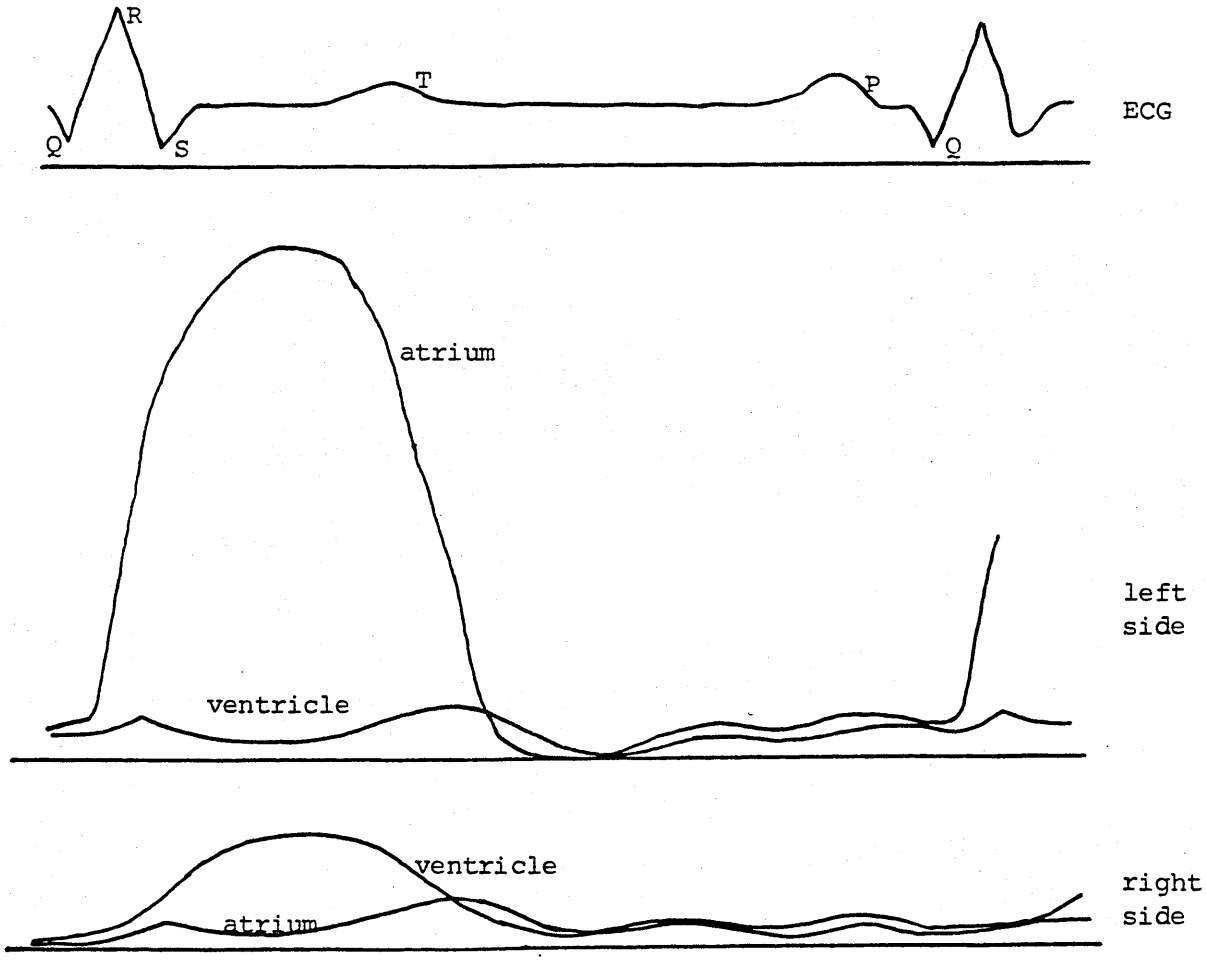


figure 3.2

The Electrocardiogram and corresponding pressure measurements on the left and right sides of the heart.

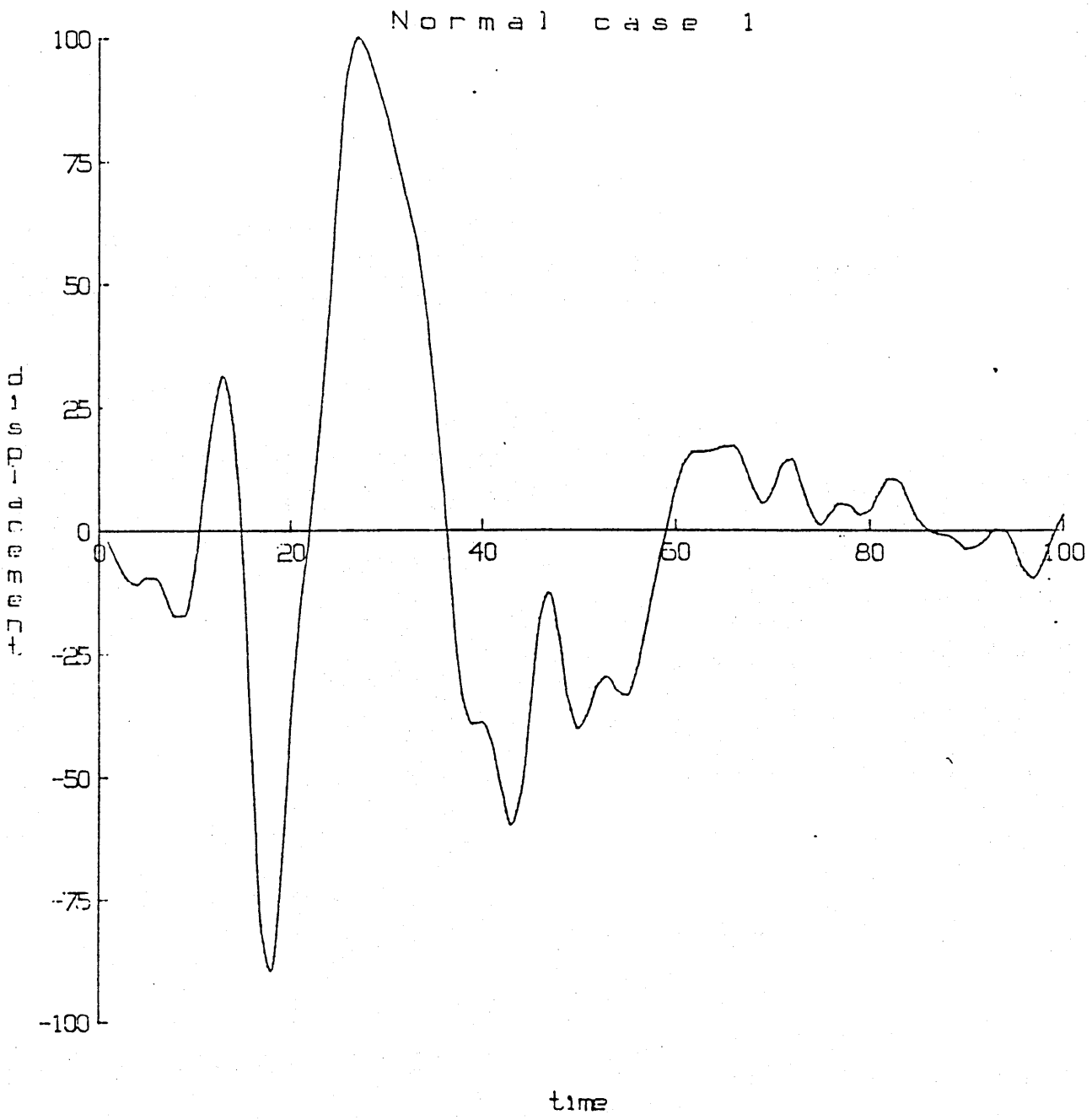


figure 3.3

An example of a healthy ballistrocardiogram

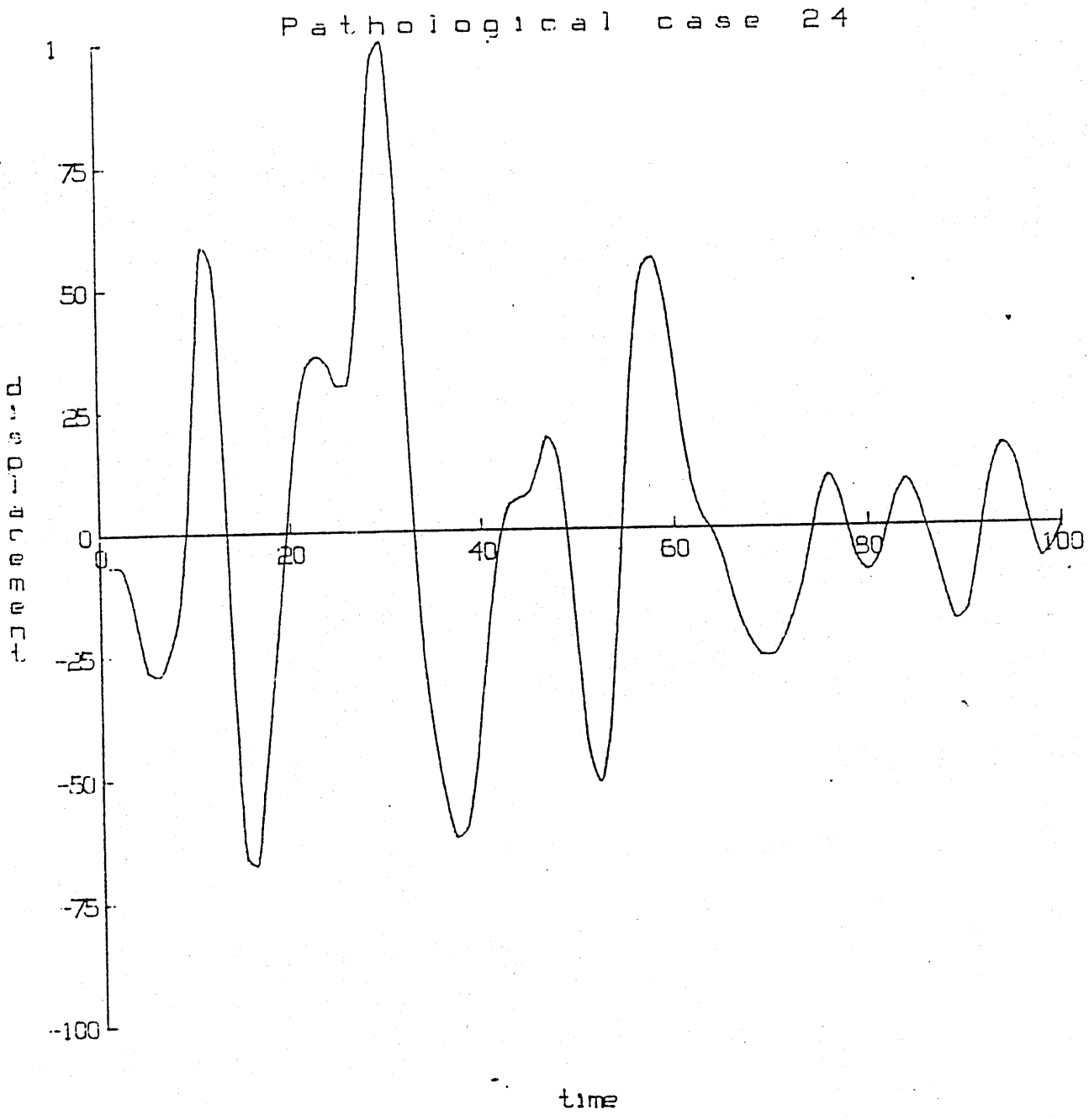


figure 3.4

An example of a pathological ballistocardiogram

3.4 Data Set 3: The Liver Scans

The idea behind the use of radionuclides in medicine is a simple one, even if in practice there are many problems. A radionuclide is a radioactive isotope, that is a substance that has the property of giving off radiation. Consequently if such an isotope is introduced into the body it can be traced by following the radiation that it produces. The problems are then to direct the radionuclide to the site of interest, to design a suitable detector, to find radionuclides that decay sufficiently quickly that they do not adversely effect the patient, and finally to interpret the pattern of radiation given off.

Since radioactive isotopes retain the same chemical properties as their non-radioactive counterparts, it is usually only necessary to find a chemical that is naturally transmitted to the site of interest. In the study of livers this problem is particularly easily solved since one of the functions of the liver is to remove unwanted substances from the blood. It is sufficient therefore to label sulphur colloid particles with any suitable isotope and then inject them into the blood stream. After about twenty minutes the particles will have been collected in the liver, although

some will also have been taken up by the spleen and bone marrow.

The radionuclide will be giving off gamma rays which can be detected by any one of a number of devices. The two main types of detector being the rectilinear scanner and the gamma camera. The only difference between the two is that the gamma camera takes a 'picture' of the liver all in one go, whilst the scanner moves backwards and forwards in a raster fashion building up the picture as it goes. Generally speaking any technique for analysing a scan produced by one method will work equally well on the other.

In the study to be used in this thesis a rectilinear scanner was used. This type of device has been used since the 1950's when it was simultaneously and independently developed in Los Angeles and Rotterdam. It consists of a lead focusing device called a collimator, a sodium iodine crystal and devices for amplication and recording. Gamma rays cannot pass through lead so that a few suitably angled holes in a lead block will act as a focusing device. A schematic representation of a scanner is shown in figure 3.5 and it will be noted that although the collimator focuses on a point, it does in fact take radiation from anywhere within its cone of view.

The effect of the gamma rays that do get through the collimator is to make a crystal scintillate, that is give off light. If the emission is amplified it may either be

stored photographically or digitised and stored as a measure of light intensity, or grey level, on a computer.

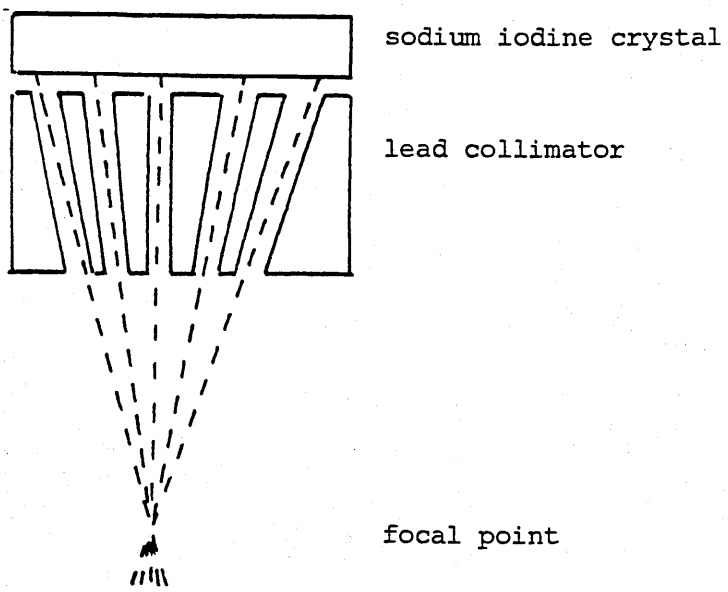


Figure 3.5

A schematic representation of a rectilinear scanner

The rectilinear scanner used in this study was an Elscint whole body scanner, model WBS-220. Like many modern scanners it has two heads that move in unison, one above the patient and one below. This means that anterior and posterior views

may be obtained in one run and afterwards the patient may be turned on their side to obtain right and left sided views.

The different isotopes that are in common use have slightly different properties. In this study one of the less powerful sources of radiation was used, namely Technetium 99m (Tc^{99m}). Consequently it tends to give a picture of what is happening in the outer regions of the liver and not at its centre. Thus an anterior view will convey little information about what is happening in the centre or at the back of the liver.

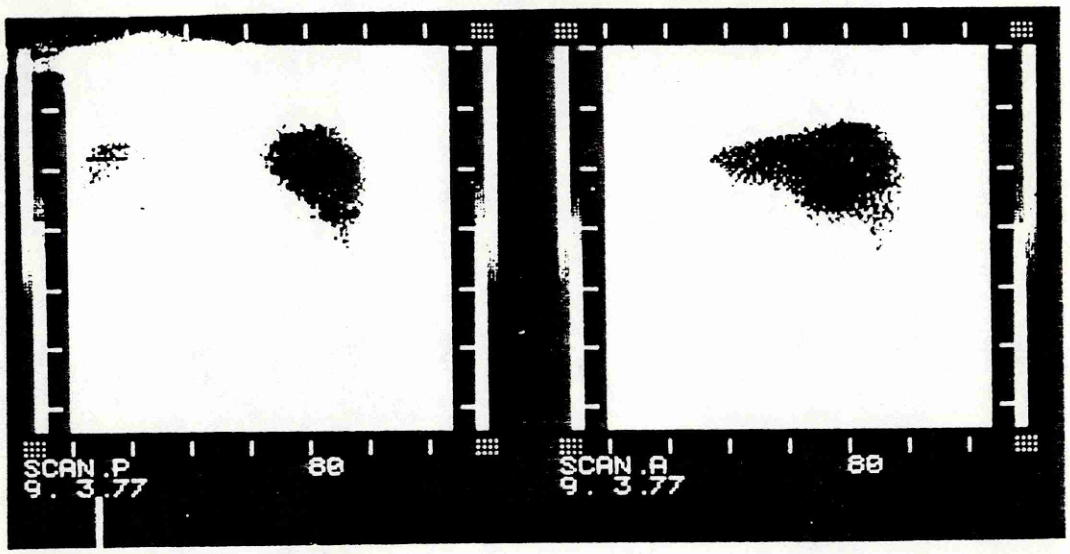
One further problem with liver scans is that they need to be positioned over the area where the liver is thought to lie. Occassionally this will mean that edges of the liver get missed and since the data was being collected routinely, it was not thought reasonable to ask for repeat scans in such cases.

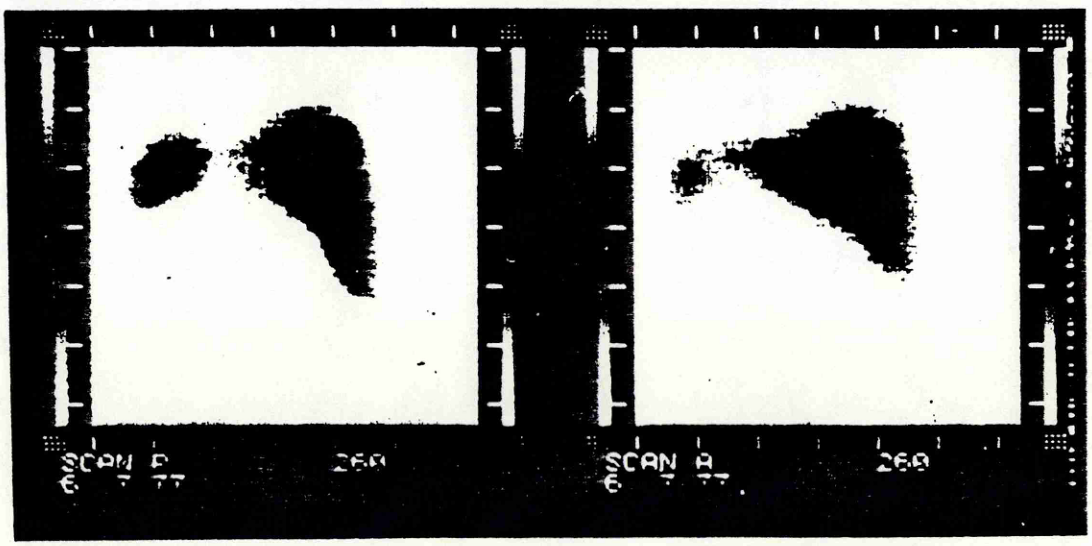
Only the anterior views have been used in this study, the data being digitised into a 128x128 grid and stored on a computer as well as being reproduced photographically. Typical photographs are shown in figure 3.6, the darker areas representing regions of high radiation.

The liver, shown diagramatically in figure 3.7, is the largest gland in the body and occupies the right and upper parts of the abdominal cavity just below the diaphram. Typically the liver weights about 1.6Kg in a healthy man and about 1.3Kg in a woman, however the variation in weights is

figure 3.6

The typical posterior and anterior liver scans





large and it may weigh as little as 1Kg or as much as 2.5Kg. The liver takes the form of two lobes, a large right and a smaller left, separated by a ligament; this division is however one of appearance rather than function.

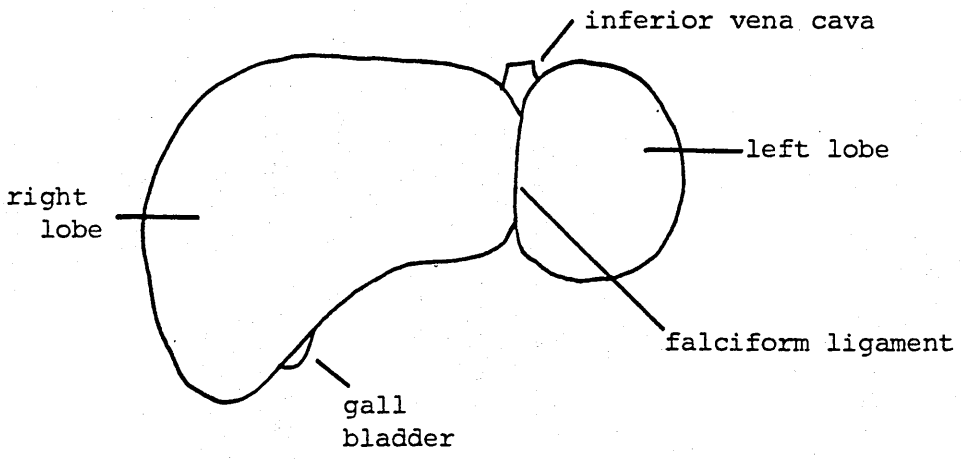


figure 3.7

A diagrammatic representation of a liver

The liver has great regenerative power and plasticity so that it can take a great variety of shapes. Its appearance will be largely dictated by the pressure of neighbouring organs although congenital variation is also possible. These natural variations can be quite marked and include the Reidel's lobe, a projection downwards from the right lobe

that may be several inches long, as well as different sizes and shapes of the two main lobes. Indeed in extreme cases the left lobe may be entirely missing, with a consequent enlargement of the right lobe. These variations are not thought to be of any diagnostic importance.

Internally the liver consists of numerous polyhedral lobules, about 1mm across, each surrounding a central vein. These lobules are responsible for a multitude of chemical reactions and are thus liable to react to many of the diseases that attack other parts of the body.

The liver's susceptibility to diseases primarily associated with other sites means that the number of causes for any abnormality in the appearance of the liver is itself large. In order to bring some sort of order to these diseases two categories have been used. The two classes, containing respectively lesions and diffuse disease, say more about the appearance of the liver when diseased rather than the causes to the illness.

Lesions are usually associated with secondary cancers, or metastases, but may also be caused by cysts or abcesses. They manifest as pockets of diseases tissue within the liver and vary in size and shape; the smallest may be barely detectable and the largest may be the size of a fist.

Diffuse diseases affect the whole liver and this may be a consequence of any number of illnesses. In its severest form the whole liver may be given over to fatty tissue, as in extreme cirrhosis, but usually the effect is much milder.

Many studies have been carried out to try to assess the worth of radionuclide imaging but the results are far from consistent. The overall percentage of patients correctly diagnosed from their liver scan has been variously quoted as being anything between 72% and 92%; Angehrn et al(1976), Braunstein and Song(1975), Conn and Elkington(1968), Jhingram et al(1971) and Vido et .al(1975). One important point to come to light in many of the studies is that there is a large variation depending upon the skill of the observer; Bland(1971), Braunstein and Song(1975), Conn and Elkington(1968) and Ludbrook et al(1972).

Biello et al(1979) performed a comparative study of radionuclide imaging and an alternative technique known as computerised tomography. Their results for the radionuclide scans are typical and show the way in which, even in the hands of experts, the scans are much better at detecting some diseases than others. Their results are summarised in table 3.1.

Biello concluded that radionuclide imaging cannot distinguish between different types of lesion, although it is good at locating them. Further it is good at detecting

Table 3.2

Results on the reading of liver scans

% correctly diagnosed from a liver scan

Diagnosis	Correct	Incorrect	Number of cases
Normal	54%	46%	37
Lesions	96%	4%	73
Jaundice	55%	45%	22
Hepatocellular	72%	28%	46
Overall	76%	24%	178

severe diffuse disease but only in cases where it could also have been detected by other clinical signs. Generally they found computed tomography to be a better technique especially in correctly identifying normal cases. Virtually the same conclusions were reached by Scherer et al(1978) after a similar but smaller study.

Clearly the percentage of cases that are correctly diagnosed will depend upon the expertise of the clinician and also on the patient mix. The more normal or mild diffuse cases that are scanned, the larger the error rate will be.

As an aid to the visualisation of the problem figures 3.8, 3.9 and 3.10 show, in three dimensions, the scans of normal and diseased livers. The heights of the plots are proportional to the readings obtained and thus the high areas correspond to the dark regions in the photographs.

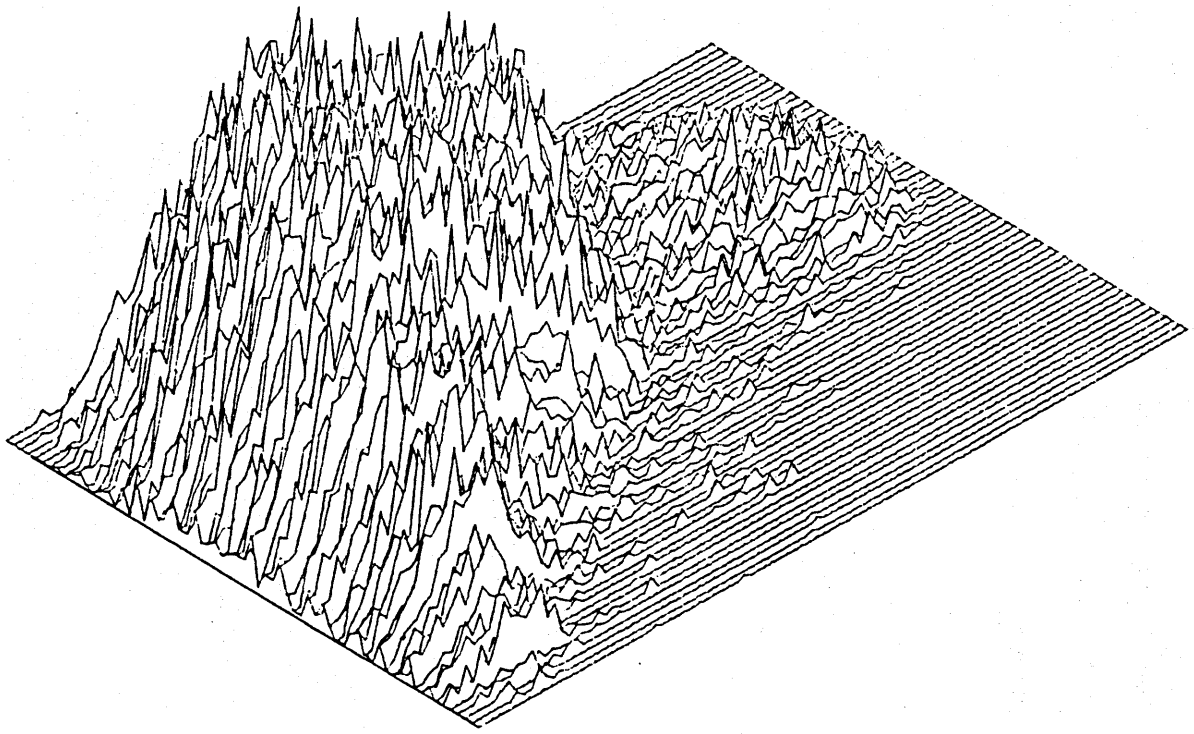


Figure 3.8

The liver scan of a normal patient

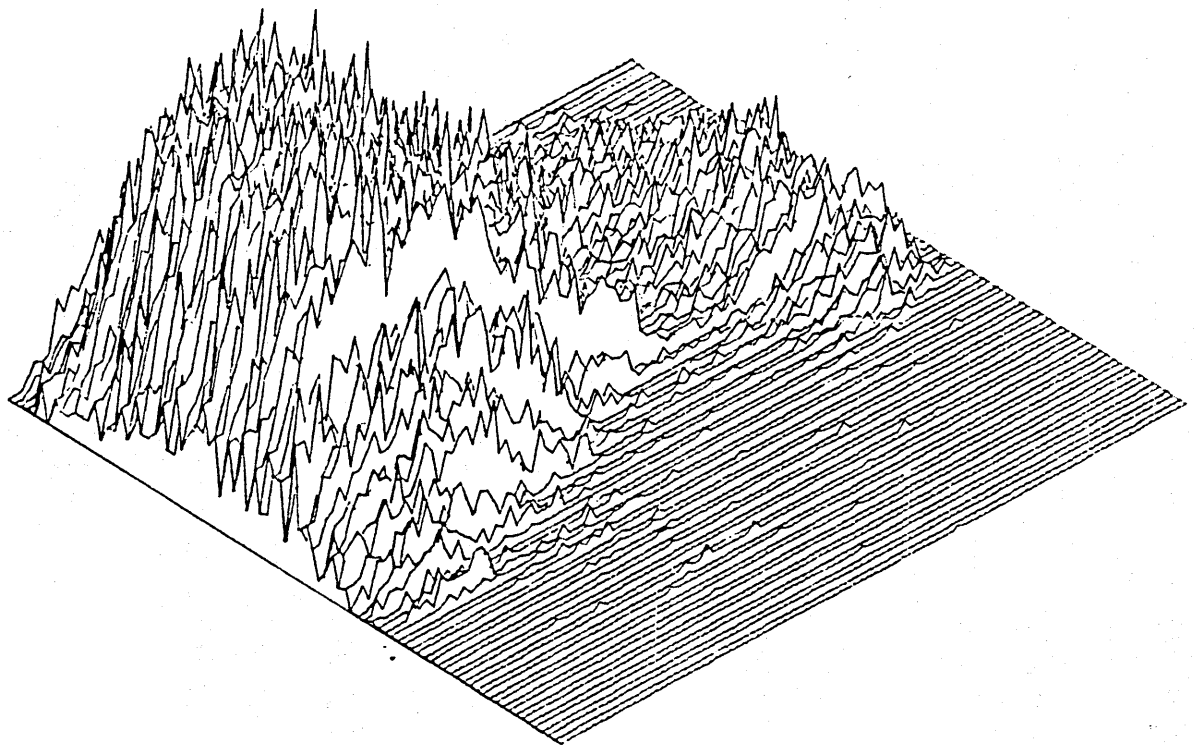


Figure 3.9

The scan of a liver affected by cancer

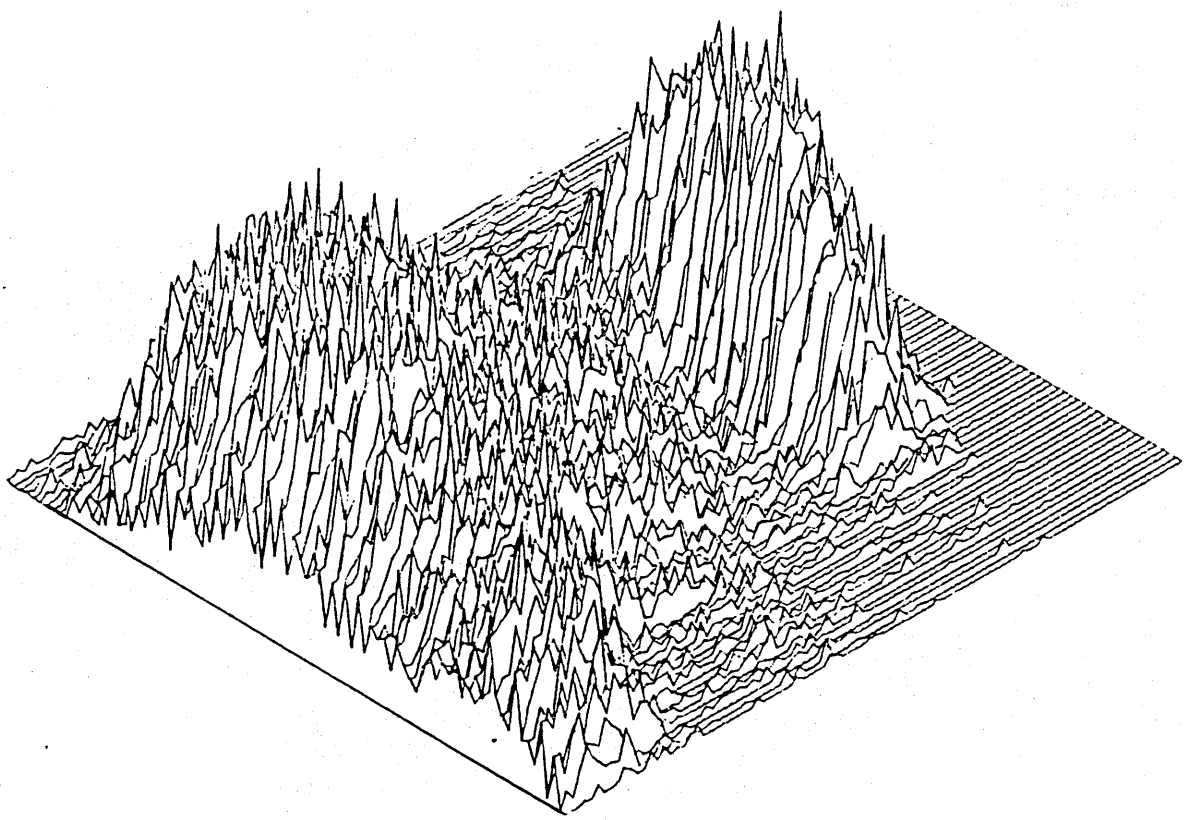


Figure 3.10

The scan of a liver affected by a diffuse disease

3.5 Randomly Generated Normal Data

In this dissertation various methods are suggested for the analysis of the three sets of medical data. As well as looking at their performance on these particular data sets it is also interesting to see how they perform under standard conditions. The nearest that we have to a standard set of conditions are classes where the data follow multivariate normal distributions and in order to obtain such data methods of random generation will be used.

Details of the many available methods of generating normal random variables can be found in such books as, Kennedy and Gentle(1980). Hoaglin and Andrews(1975) when considering the results of simulation studies make several valid points about the difficulty of interpretation due to the lack of detail that is normally provided concerning the choice of algorithm and the computer used. Following their suggestions we present a brief description of the methods employed.

The NAG library of mathematical software contains a program for generating univariate normal data using an algorithm suggested by Brent(1974). The library is implimented in double precision on a Harris 800 computer and uses a multiplicative congruential random number generator that is initialised from the computer's clock and which gives a cycle length of about 2^{57}.

Using this routine it is possible to generate standardised univariate normal variables,

$$z_i \sim N(0,1)$$

so that if we wish to generate p-dimensional normal random variables,

$$\underline{x} \sim N(\underline{\mu},\underline{\Sigma})$$

We need first to construct a p-dimensional vector of standardised normal variables,

$$\underline{z} \sim N(\underline{0},\underline{I})$$

and then use the transformation,

$$\underline{x} = \underline{A}\ \underline{z} + \underline{\mu}$$

where,

$$\underline{A}\ \underline{A}' = \underline{\Sigma}$$

A suitable factorisation of the covariance matrix may be obtained by using the Cholesky decomposition as described in, for example, Kennedy and Gentle(1980).

Other methods of generating multivariate normal data from univariate variables are possible but in a comparative study, Barr and Slezak(1972) found this method to be as good as any other.

4. DATA CHARACTERISATION

4.1 Introduction

Having introduced the three medical data sets we will now continue by considering their analysis. The first stage in any statistical analysis involves the characteristaion of the data. This effectively comprises of a pre-processing in which the original data are reconstituted into a form thought by the investigator to be a suitable starting point for the analysis.

The psychiatric data were characterised by the psychiatrist when he allocated scores to the categories and chose to amalgamate some of the symptoms. Further characterisation of this data set was found to be unnecessary although one might have tried re-scoring the categories perhaps to produce presence-absence scores.

As we saw in the last chapter, the ballistocardiograms are often summarised in terms of their peaks and troughs and in the next section we discuss the practical problems in such a characterisation. Then in sections 4.3 and 4.4 we consider a new characterisation in which the waves are modelled by a pair of damped harmonics. The coefficients of

these curves are then used as the set of potential features for discrimination, and are found to work extremely well.

The liver data present a major characterisation problem due to the size and complexity of the original set of measurements. One commonly used approach is to represent the liver in terms of the features used by clinicians in their diagnosis. Such features would include the liver's size, uptake and patchiness. In order to look at the general patterns of isotope uptake the liver scans were smoothed using a cubic spline model and it was on the basis of that fit that the features used in this study were selected. The method of fitting and the features extracted are described in sections 4.5 and 4.6 respectively.

4.2 Locating the Peaks of a Bcg.

It may be seen from the examples presented as figures 3.2 and 3.3, that the normal ballistocardiogram shows a regular damped pattern of oscillation whereas the pathological ballistocardiograms can vary considerably both from the normal pattern and from pathological subject to pathological subject. In these abnormal cases it is not possible to say with any degree of certainty just how many peaks there will be or where they will occur. Thus in the most abnormal cases the very concept of a specific peak may be doubtful.

Because of the irregularity of the abnormal ballistocardiograms some experimentation was needed before a program was produced that was robust enough to work on all cases. The algorithm finally adopted worked along the length of the wave locating maxima and minima and fitting quadratics around the turning points.

In the original search along the wave a point with amplitude y_i is accepted as a maximum if

(i) $y_i = \text{Max}(y_{i-3}, y_{i-2}, y_{i-1}, y_i, y_{i+1}, y_{i+2}, y_{i+3})$

(ii) $y_{i+1} > y_{i+2}$ or y_{i+3}

and

$y_{i-1} > y_{i-2}$ or y_{i-3}

Thus, amongst others, the situations illustrated in figure 4.1 would all be accepted as maxima. This definition was found to be necessary since, despite the averaging and filtering that had already been performed during the pre-processing of the waves, they were still subject to random variations. A similar definition was used for the minima.

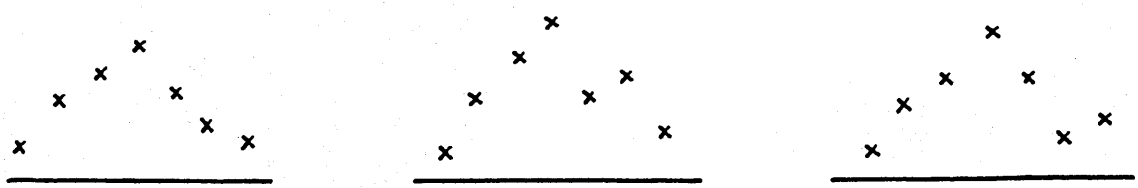


figure 4.1

Configurations that would be acceptable as maxima

Having located the turning point a quadratic was fitted through that point and three points on either side of it. The maximum or minimum of the quadratic was then taken as the peak or trough. Both the amplitude and location of the fitted turning points were recorded. In this way the program found the location and amplitude of the H,I,J,K,L,M and N waves giving a set of 14 potential features.

4.3 Modelling the Ballistocardiograms: Background

The model that will be used for the ballistocardiograms consists of super-imposed damped harmonic waves of the form,

$$A \exp(-bt) \sin(c+dt)$$

The practical problems involved in such a fit, together with a consideration of the results, may be found in the next section. Here we are concerned only with the statistical background to the fitting of such models.

The method of fitting that was used is the method of maximum likelihood. This approach rests on the assumption that, given a statistical model, all of the information which the data can provide is contained in the likelihood. Following the assertion it is natural to choose values for the parameters of the model that make the likelihood as large as possible. Thus if $p(\underline{X} | \Psi)$ is a model for the density of the variable $\underline{X}$ containing unknown parameters then we would estimate Ψ by selecting the values for which the likelihood,

$$L = \prod_{i=1}^{n} p(\underline{x}_i \mid \Psi)$$

is maximised. Here $\underline{x}_i$ represents a values from a random

sample of data.

The estimates produced by this method may well be biased and inefficient but as demonstrated in, for example, Silvey(1970), when the samples are large the bias will be small and the variances of the estimates will approach their Cramer-Rao bound. Thus the variance-covariance matrix of the estimates is approximated by,

$$-E\left(\frac{\partial^2 \ln(L)}{\partial\Psi_i \partial\Psi_j}\right)_{\hat{\underline{\Psi}}} = E\left(\frac{\partial \ln(L)}{\partial\Psi_i} \quad \frac{\partial \ln(L)}{\partial\Psi_j}\right)_{\hat{\underline{\Psi}}}$$

In our particular example we have variables $\underline{Y}$ representing the amplitude of the wave and $\underline{X}$ representing the time through the cycle. They are to be related by the model,

$$\underline{Y} = f(\underline{X} \mid \underline{\xi}) + \underline{\varepsilon}$$

where f is a non-linear function of the parameters $\underline{\xi}$. Further we suppose that the random component $\underline{\varepsilon}$ has a distribution ,

$$p(\underline{\varepsilon}|\underline{\Psi})$$

which itself depends upon unknown parameters $\underline{\Psi}$. Now given a set of points y_i and x_i, i=1,...,n we may find and maximise the likelihood,

$$L = p(\underline{\varepsilon} \mid \underline{\Psi})$$
$$= p(\underline{y} - f(\underline{x} \mid \underline{\varepsilon}) \mid \underline{\Psi})$$

Usually with the exponential family of densities it is

simpler to minimise the quantity $-\ln(L)$ rather than maximise L but these are obviously equivalent.

This approach offers great flexibility and standard software exists for fitting such models with a variety of error structures. However they generally require, for ease of computation that the models be linear in the parameters. In our present case this is not so and it is necessary to optimise the likelihood using a specially prepared program employing one of the many methods for non-linear optimisation.

The most frequently used method of optimisation is the so called Newton's method. Suppose that $\theta' = (\underline{\xi}', \underline{\Psi}')$ is the set of parameters that we wish to estimate and that $\Phi = -\ln(L)$ is the function that we wish to minimise. If $\underline{g}$ is the vector of first derivatives of Φ and $\underline{H}$ is the Hessian matrix of second derivatives, that is,

$$g_i = \frac{\partial \Phi}{\partial \theta_i} \qquad H_{ij} = \frac{\partial^2 \Phi}{\partial \theta_i \partial \theta_j}$$

Then close to the minimum we might hope to approximate Φ by Q, its Taylor series truncated at the quadratic term. Thus if $\underline{\theta}_o$ is a particular point in the region of the minimum,

$$Q(\underline{\theta}) = \Phi(\underline{\theta}_o) + \underline{g}'(\underline{\theta}_o)(\underline{\theta} - \underline{\theta}_o) + \tfrac{1}{2}(\underline{\theta} - \underline{\theta}_o)' \underline{H}(\underline{\theta}_o)(\underline{\theta} - \underline{\theta}_o)$$

The minimum of Q is easily found since,

$$\frac{\partial Q}{\partial \underline{\theta}} = \underline{g}(\underline{\theta}_o) + \underline{H}(\underline{\theta}_o)\ (\underline{\theta} - \underline{\theta}_o)$$

which when equated to zero gives, providing that $\underline{H}$ is non-singular,

$$\underline{\theta} = \underline{\theta}_o + \underline{H}^{-1}\ (\underline{\theta}_o)\ \underline{g}(\underline{\theta}_o)$$

Consequently we may take any $\underline{\theta}_o$ and locate the minimum of Q in a single step. For most functions Φ the minimum will not coincide exactly with the minimum of Q but we should at least be closer to the desired minimum. This then suggests the use of the iterative process known as Newton's method,

$$\underline{\theta}_{i+1} = \underline{\theta}_i + \underline{H}^{-1}\ (\underline{\theta}_i)\ \underline{g}(\underline{\theta}_i)$$

4.4 Modelling the Bcg's:Practical Considerations & Results

Non-linear optimisation is fraught with many practical difficulties. The function -ln(L) described in section 4.3 is prone to having multiple minima and it is necessary to try several starting points before one can be reasonably sure that the global minimum has been reached. Further the usual form of Newton's method requires a matrix inversion; in practice one often finds that at some stage in the fit the Hessian matrix becomes singular. To overcome these and other practical problems several refinements were incorporated into the fitting program.

Firstly the matrix $\underline{H}$ was replaced by an approximation $\underline{N}$ that is easier to compute and non-singular over a wider range of parameter values. This is the so called, Method of Gauss, described in, for example, Bard(1974). Occassionally it is found that even $\underline{N}$ will be singular in which case it is in turn replaced by $\delta\underline{I}$ where δ is some small constant and $\underline{I}$ is the identity matrix. Thus when Gauss's method is not possible we resort to the much slower method of steepest descent.

With some functions it is possible that the step suggested by the quadratic approximation actually overshoots

the true minimum and produces a larger function value. That is that,

$$\Phi(\underline{\theta}_{i+1}) > \Phi(\underline{\theta}_i)$$

To avoid following such steps the method of reduced gradient was employed. This involves checking on the function value at each stage and, if it is not an improvement, one takes a point part way along the line joining $\underline{\theta}_i$ and $\underline{\theta}_{i+1}$. Thus instead of,

$$\underline{\theta}_{i+1} = \underline{\theta}_i - \underline{H}^{-1}(\underline{\theta}_i)\, \underline{g}(\underline{\theta}_i)$$

we take,

$$\underline{\theta}_{i+1} = \underline{\theta}_i - a\, \underline{H}^{-1}(\underline{\theta}_i)\, \underline{g}(\underline{\theta}_i) \qquad 0 < a < 1$$

Finally there is the numerical problem due to the finite word length of the computer. Occasionaly wild steps may lead to parameter values that cause the computer to overflow or underflow. To avoid this each parameter was constrained to lie between two values l_j and u_j.

Then if,

$$\theta_{i+1,j} > u_j$$

it is replaced by,

$$\theta_{i+1,j} = .75\, u_j + .25\, \theta_{i,j}$$

or if,

$$\theta_{i+1,j} < l_j$$

it is replaced by,

$$\theta_{i+1,j} = .75\, l_j + .25\, \theta_{i,j}$$

With these precautions it was found that even with a relatively crude initial guess the minimum could be located to five significant figure accuracy within about twenty steps.

The appearance of the ballistocardiograms suggests a pair of superimposed damped harmonic waves starting some way after the beginning of the peak of the QRS wave of the electrocardiogram. Thus the model used for the systematic variation was,

$$f_t = \mu \qquad t < t_o$$

$$f_t = \mu + a_1 e^{-b_1 t} \sin(c_1 + d_1 t) + a_2 e^{-b_2 t} \sin(c_2 + d_2 t) \qquad t \geq t_o$$

The best value for t_o varies from case to case but twelve is a useful approximation for all cases.

One would expect that the error structure for this model would be very complex since the residuals are certainly not going to be independent but will display a high degree of autocorrelation. However, following the suggestions of Box and Jenkins(1976), we will make an initial assumption of normal errors with constant variance. Thus,

$$L = \prod_{t=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}e_t^2/\sigma^2}$$

and

$$-\ln(L) = 50\ln(\sigma^2) + 50\ln(2\pi) + \frac{1}{2\sigma^2}\sum_{t=1}^{100} e_t^2$$

Differentiating with respect to σ^2 and equating to zero we obtain the usual maximum likelihood estimator of σ^2, namely

$$\hat{\sigma}^2 = 0.01 \sum_{t=1}^{100} e_t^2$$

and see also that $-\ln(L)$ is minimised when Σe_t^2 is minimised. Thus in this case maximum likelihood reduces to least squares. Experience suggests that whilst the use of the wrong error structure will not greatly effect the parameter estimates it is likely to cause us to underestimate their standard errors. This does not matter so much in our application since we only want the parameter estimates as features for a discriminant analysis.

In the case we have described Gauss's method has a particularly simple form. The function to be minimised is,

$$\Phi = \sum_{t=1}^{100} e_t^2 = \sum_{t=1}^{100} (y_t - f_t)^2$$

so that the first derivatives are,

$$\frac{\partial \Phi}{\partial \theta_i} = 2 \sum_{t=1}^{100} e_t \frac{\partial e_t}{\partial \theta_i} = -2 \sum_{t=1}^{100} e_t \frac{\partial f_t}{\partial \theta_i}$$

and the second derivatives are,

$$H_{ij} = \frac{\partial^2 \Phi}{\partial \theta_i \partial \theta_j} = 2 \sum_{t=1}^{100} \frac{\partial f_t}{\partial \theta_i} \frac{\partial f_t}{\partial \theta_j} - 2 \sum_{t=1}^{100} e_t \frac{\partial^2 f_t}{\partial \theta_i \partial \theta_j}$$

The approximation $\underline{N}$ is obtained by using the first of these two terms, thus only the first derivatives need be evaluated.

The necessary first derivatives are easily obtained by t he usual analytic means,

if $t < t_o$

$$\frac{\partial f_t}{\partial \mu} = 1 \qquad \frac{\partial f_t}{\partial \theta} = 0 \qquad \theta = (a_1, b, \ldots, d_2)$$

if $t \geq t_o$

$$\frac{\partial f_t}{\partial \mu} = 1$$

$$\frac{\partial f_t}{\partial a_i} = e^{-b_i t} \sin(c_i + d_i t)$$

$$\frac{\partial f_t}{\partial b_i} = -a_i\, t\, e^{-b_i t} \sin(c_i + d_i t)$$

$$\frac{\partial f_t}{\partial c_i} = a_i\, e^{-b_i t} \cos(c_i + d_i t)$$

$$\frac{\partial f_t}{\partial d_i} = a_i\, t\, e^{-b_i t} \cos(c_i + d_i t)$$

$i = 1,2$

Table 4.1 shows a summary of the results obtained when the model was fitted to the 81 normal cases with $t_o = 12$. In all cases fitting was quick even from a crude initial guess.

Table 4.1

Summary of fit for Normal cases

Parameter	Mean	Standard deviation	Minimum	Maximum
μ	-1.74	.97	-3.77	1.17
a_1	145	48	90	278
b_1	.032	.007	.007	.046
c_1	3.12	.37	1.73	3.72
d_1	.17	.02	.13	.21
a_2	254	98	111	508
b_2	.059	.014	.031	.092
c_2	3.46	.34	2.76	4.31
d_2	.37	.04	.28	.44
σ^2	101	37	46	287

The abnormal ballistocardiograms, being irregular produced more of a problem for the fitting program. Indeed the last seven of the fifty cases, presumably repeat measurements from the same individual, showed no damping at all; on the contrary the largest peaks occur at the end of the cycle. The model can be made to fit in such cases but the fit is so poor as to be worthless. Since we are fitting

with a view to later discrimination the most sensible policy seems to be to omit these obviously non-normal cases. The summary for the remaining 43 abnormal cases is contained in table 4.2.

Table 4.2

Summary of fit for 43 abnormal cases

Parameter	Mean	Standard deviation	Minimum	maximum
μ	-1.74	.97	-3.77	1.17
a_1	145	48	90	278
b_1	.026	.010	.009	.063
c_1	2.89	.60	.61	3.92
d_1	.180	.025	.130	.230
a_2	137	82	38	543
b_2	.026	.016	.000	.081
c_2	3.22	.62	2.23	5.03
d_2	.36	.045	.25	.44
σ^2	392	95	201	635

Clearly the basic difference between the two classes lies in the degree of damping. Typically the normal waves have a larger amplitude and larger damping factor than the abnormals. This is well illustrated in figure 4.2 which shows the curves obtained by using the average values of the parameters for each class. The lack of damping is clearly

evident.

Although the shapes of the fitted curves are highly diagnostic the variablity about the fit, as measured by the variance, also contains much useful discriminatory information; there being roughly four times the variance about the model in the abnormal cases. If we compare the sums of squares about the model with the sums of squares about the mean in the usual manner for regression analysis one finds that the normal fits explain about 95% of the variablity, whilst the abnormal fits explain about 90%.

As an illustration of the quality of the fit, figure 4.3 shows the fitted curve and original ballistocardiogram for the first normal case. This case is typical of the fits produced and the plot of residuals, figure 4.4, shows that the error is far from random. In fact we can see from the autocorrelations, shown in figure 4.5, that there is still a fairly major cycle of period 12 that has not been picked up by this model.

These plots suggest better models that might be used to explain the ballistocardiograms. Obvious improvements include,

(a) using more than two waves,

(b) allowing t_o to vary,

(c) allowing the superimposed waves to start at different points,

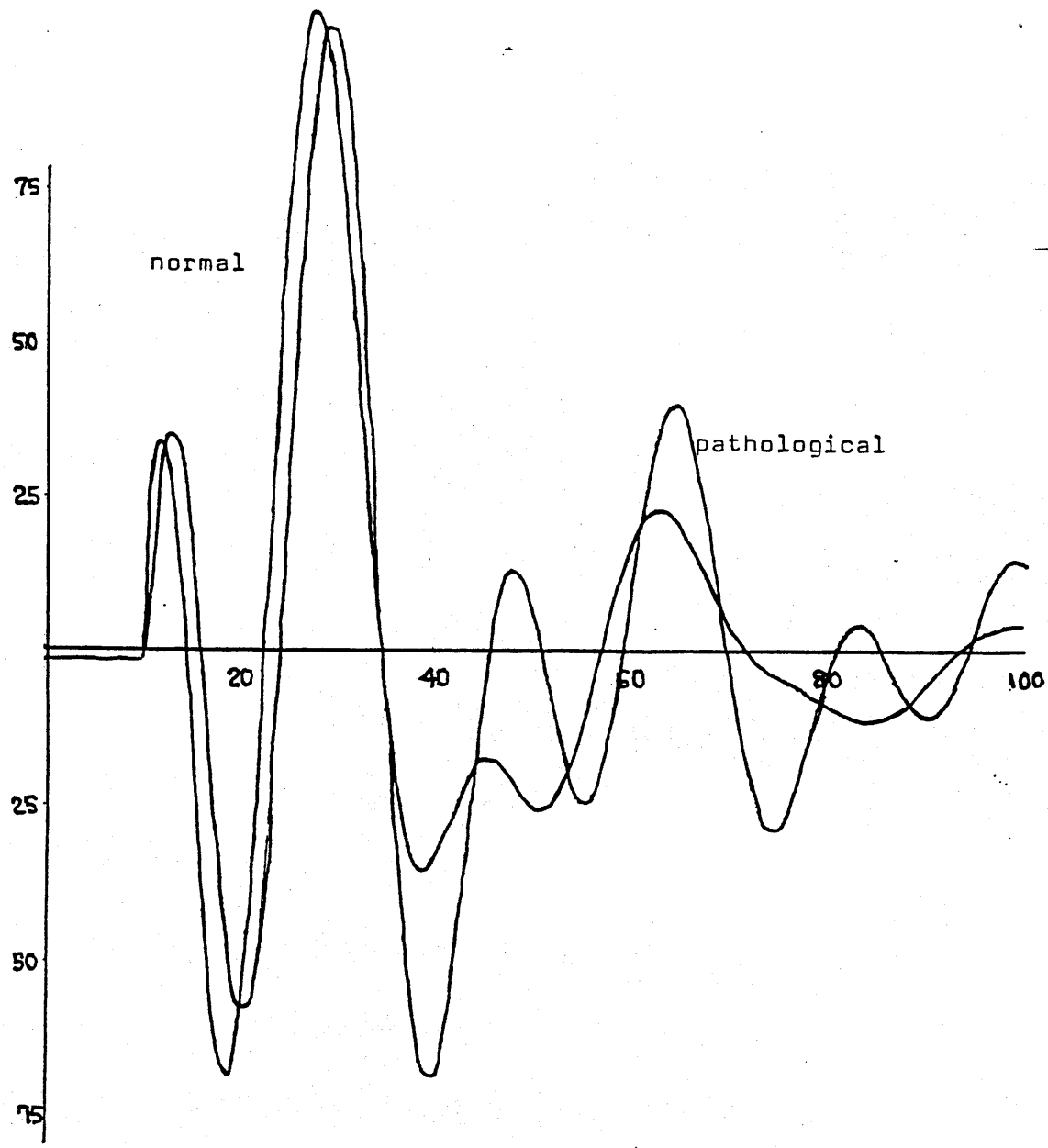


Figure 4.2

Average fits for normal and pathological ballistocardiograms

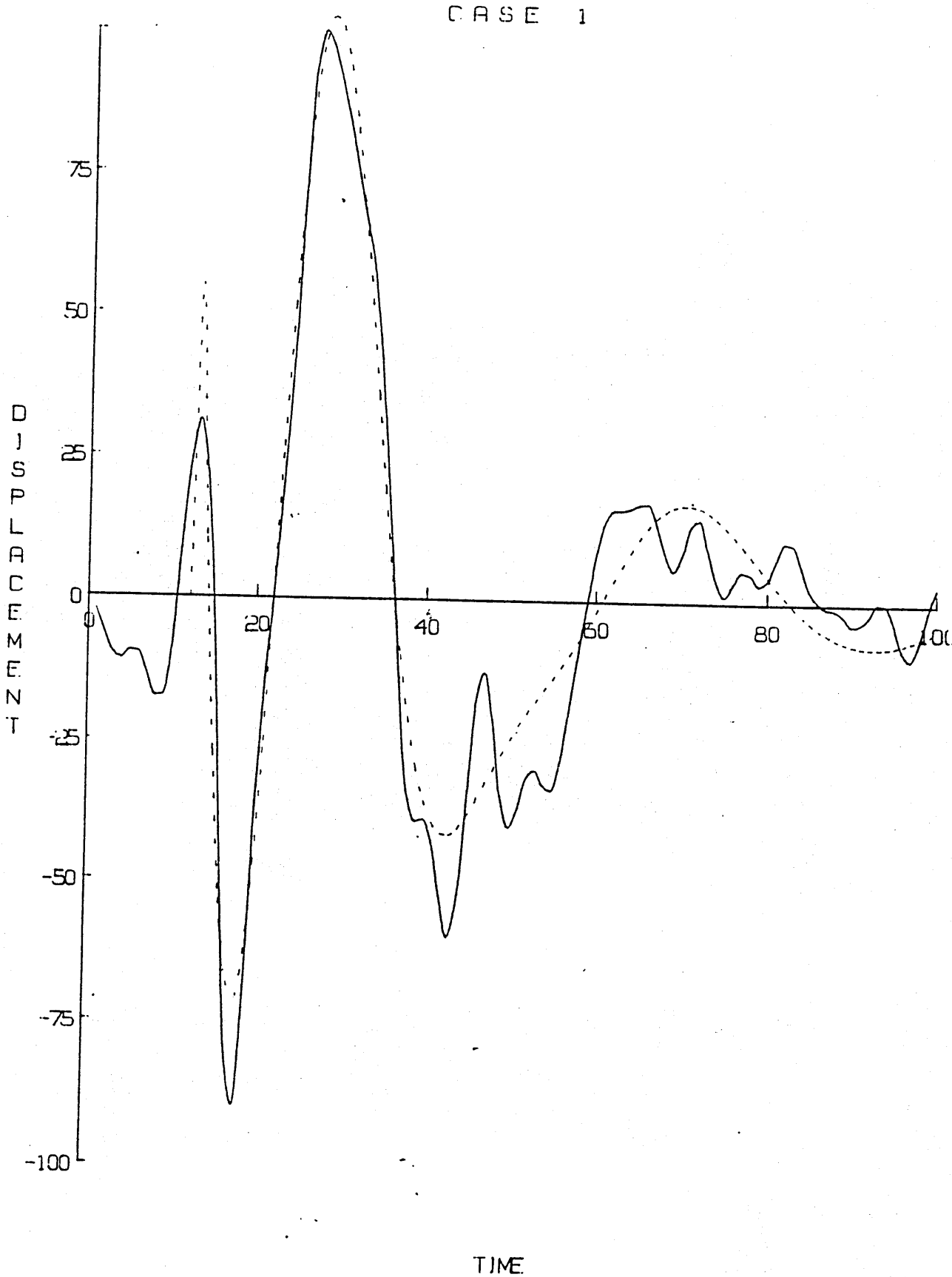


Figure 4.3

An example of a healthy Bcg and its fitted curve.

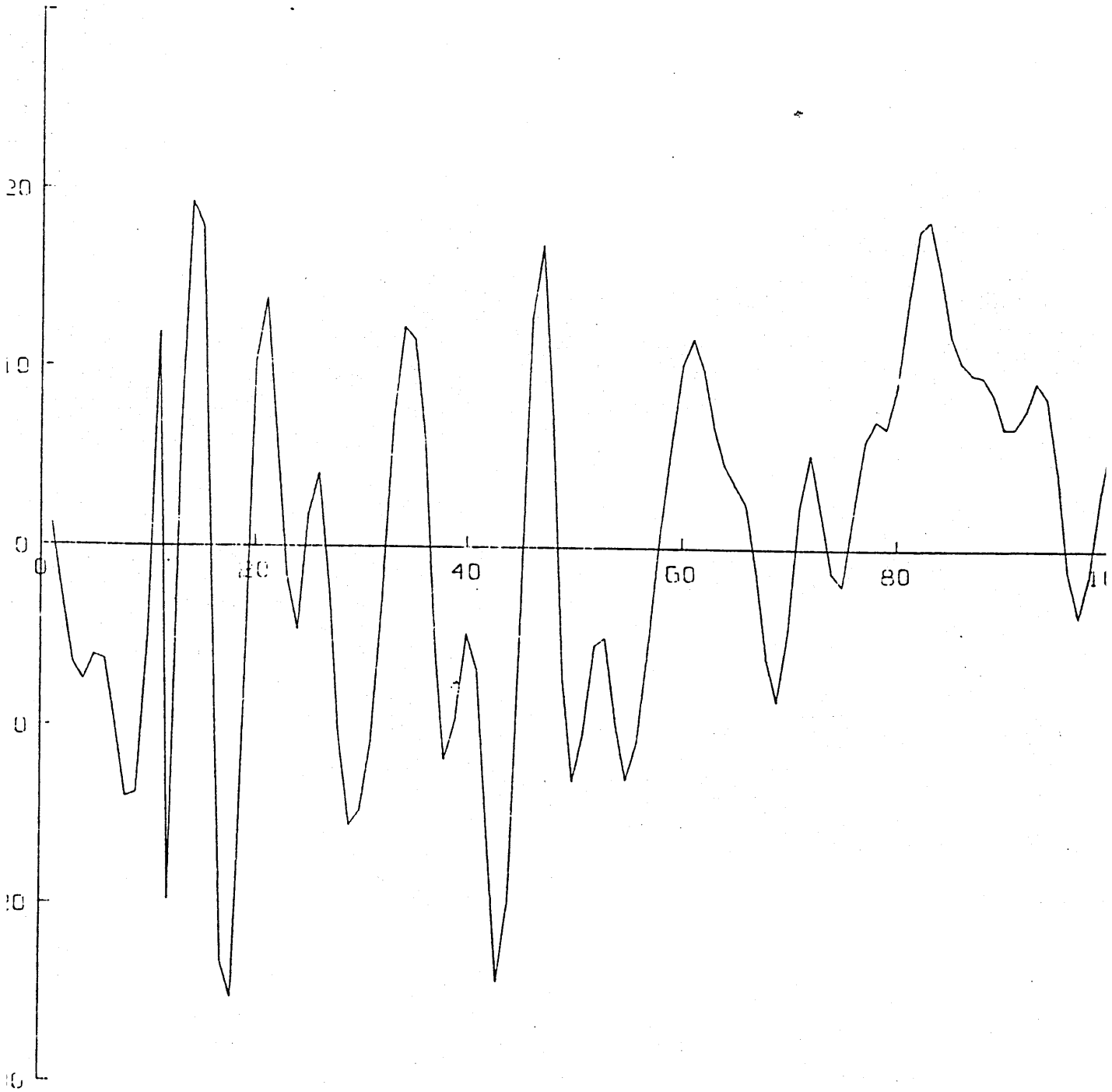


Figure 4.4

The residuals from the fit shown in Figure 4.3

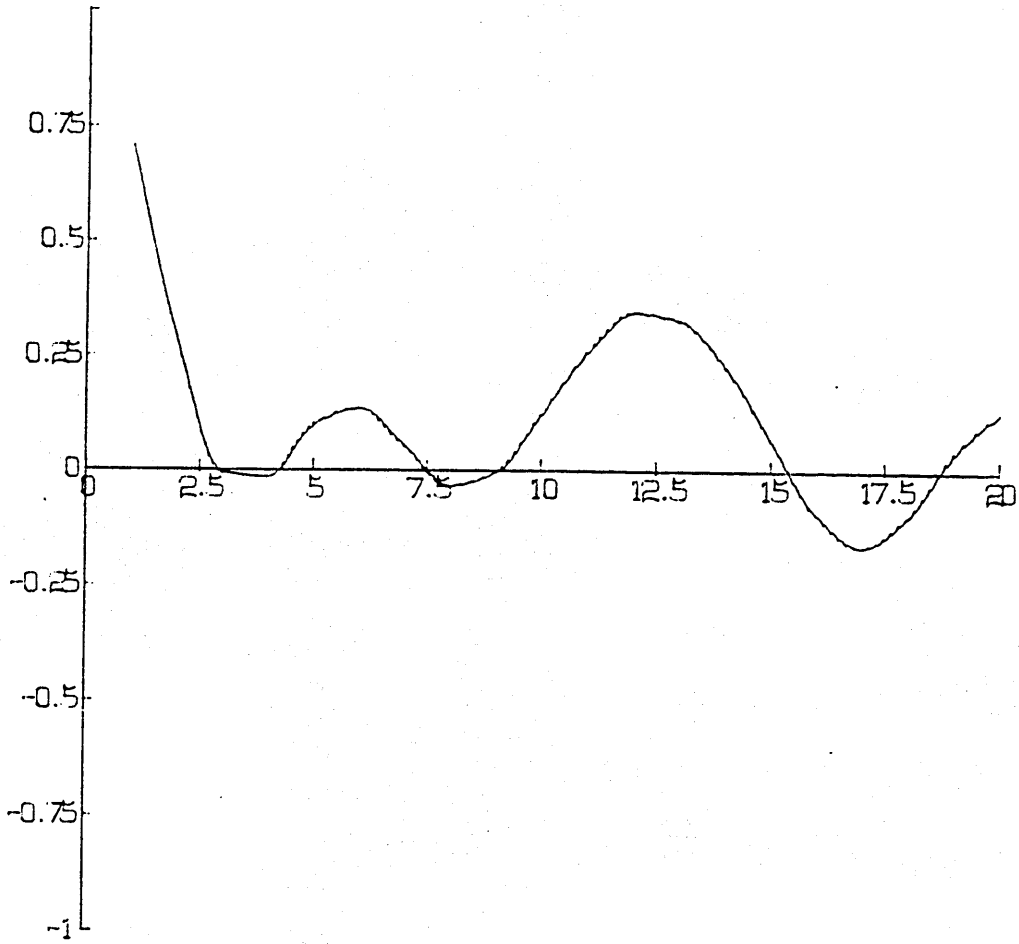


Figure 4.5

The autocorrelations between the residuals from the fit shown in figure 4.3

(d) wrapping the waves around so that the end of one cycle joins onto the beginning of the next,

(e) finding a more realistic model for the error structure.

These possibilities were not pursued at this stage for two reasons. Firstly we have already achieved our objective of characterising the waves for subsequent feature selection and secondly it would be wrong to press on with more elaborate models without the close co-operation of a medical expert who could give an interpretation to the results.

4.5 Image Characteristation: Background

In a recent review Rosenfeld(1984) considered the latest developments in image analysis and showed how the subject has progressed by using a large number of ad hoc procedures. It is because of this largely problem dependent nature of image analysis that we have chosen to distinguish the characteristaion of the image from the more general methods of feature selection.

When analysing the liver scans we wish to pick out a number of specific clinical features. Physicians generally use the liver scans to produce an overall description of the liver and to locate regions of low uptake. In attempting an overall description we are drawn into the fields of image enhancement, edge detection, texture and shape measurement, whilst the location of lesions is a problem in segmentation. The main difficulty with the lesions is that they do not take predictable shapes and so the commonly used techniques for template matching are not applicable.

Despite the problem dependent development of image analysis two main approaches can be discerned, namely the use of moving operators and the use of the two dimensional Fourier transform. Both of these techniques can be modified

to emphasise some features of the image and play down others.

Suppose that we represent the scan by an nxn matrix of grey scale scores,

$$g(x,y) \qquad x,y=1,\ldots,n$$

Then a moving 3x3 operator would consist of weights,

$$w(i,j) \qquad i,j=-1,0,1$$

used to transform the old image g into a new image h, where,

$$h(x,y) = \frac{\sum_{ij} w(i,j)g(x+i, y+j)}{\sum_{ij} w(i,j)}$$

Figure 4.6 shows a commonly used 3x3 operator, namely the simple moving average, useful for smoothing. Other windows such as the Laplacian, that is the sum of the two second partial derivatives, help to enhance contrasts. Numerous variations on this theme have been used; notably these include the use of windows of different sizes and the use of logical operators. The later group involve a decision as to the form of the transformation to be used which will be based on the values found within the window.

The Fourier transform involves a conversion of the original image into a form based on spacial frequencies u,v.

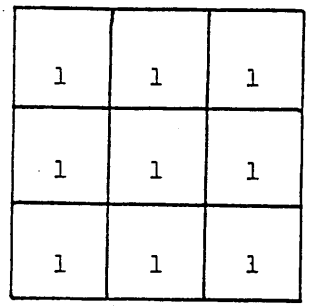


Figure 4.6

The moving average window

Thus the Fourier transform of g(x,y) is defined as G(u,v) where,

$$G(u,v) = \frac{1}{N} \sum_{x,y=1}^{N} g(x,y) \exp\left[\frac{-2\pi i}{N} (ux + vy)\right]$$

Having obtained the frequency representation this can be transformed by a suitable chosen function W(u,v) so that the transformed frequency representation H(u,v) is obtained,

$$H(u,v) = G(u,v)\ W(u,v)$$

The function W can be chosen to preserve selected frequencies at the expense of others and after it has been used we may return to the original representation by using the inverse Fourier transform,

$$h(x,y) = \frac{1}{N} \sum_{u,v=1}^{N} H(u,v) \exp\left[\frac{2\pi i}{N} (ux + vy)\right]$$

This whole process is the two dimensional generalisation of the method of preprocessing used on the ballistocardiograms.

4.6 Image Characterisation: Pre-processing

Before the data were made availablethey were subjected to a crude and irreversible pre-processing. The value seven had been subtracted from each reading in the image, resulting negative values being replaced by zero. This value had been arrived at by trial and error and was supposed to have removed the noise due to the isotope in the body that had not collected in the liver.

In fact an inspection of the scans shows that this pre-processing had only been a partial success for the livers were still not clearly defined. Consequently a second pre-processing was necessary. This second stage consisted of logically smoothing the scan so that if fewer than two of the four immediate neighbouring points were non-zero then the point was replaced by zero and if all four neighbours of a zero point were non-zero, then that point was replaced by 0.5. Otherwise the point was left unchanged. Thus,

if n_o = number of non-zero values amongst

(g(x+1,y),g(x-1,y),g(x,y+1),g(x,y-1))

then

if $n_o > 2$ $h(x,y)=g(x,y)$

if $n_o < 2$ $h(x,y)=0$

if $n_o=4$ and $g(x,y)=0$ $h(x,y)=.5$

This filtering was repeated four times, experiment showing that this indeed remove the background noise and hence leave a clearly defined liver. Figures 4.7 and 4.8 show the effect of this pre-processing on one of the scans. The lack of definition in the original is clearly shown as is the clarity of the final version. Both figures are scaled so that,

g(x,y)	symbol
0.5 - 7	1
8 - 15	2
:	:
:	:
64 - 71	9
72 - 79	A
:	:

Figure 4.7

A Typical Scan before the second Pre-processing

Figure 4.8

The same scan after the second Pre-processing

4.7 Liver Scan Modelling

In order to select the features of interest from the liver scans it is first necessary to obtain a smoothed representation of the scan's surface. This could be produced by the use of a moving operator or by using the Fourier transform but for special reasons associated with the nature of the data a different type of smoothing was used.

Inspection of the normal scans shows that they typically take the form of two bulges corresponding to the two lobes. Abnormalities show up as deviations from this overall pattern. Typically a diffuse disease shows up as a overall lower uptake and a patchy distribution of the isotope, whereas lesions appear as distinct regions of low uptake. The ideal way to locate such lesions would be to superimpose on the scan a reconstruction of the normal pattern of the same liver without those lesions. Any type of smoothing will to a degree achieve this but because of the varieties in lesion size it is difficult to avoid allowing the smooth to follow the undulation of the lesion. For this reason it was decided that a more rigid structure needed to be imposed on the scans.

The rigid structure used was in the form of a cubic

spline surface fitted to the scan. The surface was given sufficient freedom to allow it to rise and fall with the bulges of the liver but not enough to allow it to follow the smaller variations.

Spline fitting is most easily explained in one dimension and then expanded to two dimensions by analogy. Suppose that we were to fit a spline curve through a set of data points,

$$(x_i, f_i) \quad i=1,\ldots,n$$

We would first need to divide the range of the x values, denoted by (a,b), into a series of intervals by defining points $\lambda_1, \lambda_2, \ldots, \lambda_h$, known as knots. One then has the situation shown in figure 4.9.

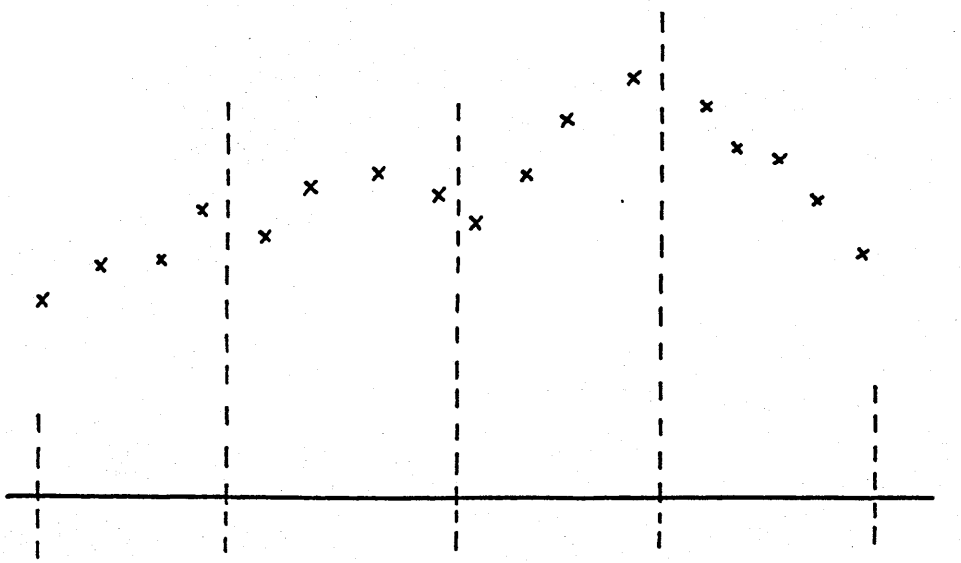


Figure 4.9

A data set divided by 5 knots

A cubic spline fit to these data would consist of cubics fitted to each of the intervals in such a way that they join at their ends and are continuous in their first two derivatives. Greville(1968) showed that such a fit could be represented in the form,

$$S(x) = \sum_{j=0}^{3} a_j x^j + \sum_{i=1}^{h} b_i (x - \lambda_i)_+^3$$

where,

$$E_+ = E \quad \text{if } E>0$$
$$= 0 \quad \text{if } E<0$$

Fitting a cubic spline s(x) by least squares becomes a problem in minimising,

$$\sum_{i=1}^{n} (f_i - s(x_i))^2$$

Unfortunately although this is the simplest representation of the problem it is computationally unsound as the determination of the parameters is likely to be ill-conditioned. For this reason the cubic B-spline was suggested.

A B-spline $M_j(x)$ is a cubic spline defined over four consecutive intervals $\lambda_{j-4}, \ldots, \lambda_j$ which is zero everywhere

else. To obtain full coverage of the range (a,b) this method will therefore require the specification of extra knots outside (a,b). Namely,

$$\lambda_{-3} < \lambda_{-2} < \lambda_{-1} < \lambda_{o} < a$$

$$b > \lambda_{h+1} > \lambda_{h+2} > \lambda_{h+3} > \lambda_{h+4}$$

The actual values chosen for these extra knots do not effect the solution.

Within the range (a,b) the fitted curve is then given by,

$$S(x) = \sum_{j=1}^{h+4} c_j M_j(x)$$

This numerically stable fitting procedure can now be extended into two dimensions by supposing that we have to fit data of the form,

$$(x_i, y_i, f_i) \qquad i=1,\ldots,n$$

and that the ranges of values of x and y are divided up by knots as represented in figure 4.10.

The form of the fitted surface is then,

$$S(x,y) = \sum_{j=1}^{h+4} \sum_{k=1}^{h+4} c_{jk} M_j(x) N_k(y)$$

where M_j and N_k are B-cubic splines defined over sets of

intervals for x and y respectively. Full computational details may be found in Hayes and Haliday(1974) and the algorithm has been implimented as a part of the NAG library.

Figure 4.11 illustrates the degree of fit obtainable by such cubic surfaces, by showing the fit obtained when the procedure is applied to the scan shown in figure 3.7. The points outside the region of the liver are given zero weight during the fit and assumed zero afterwards.

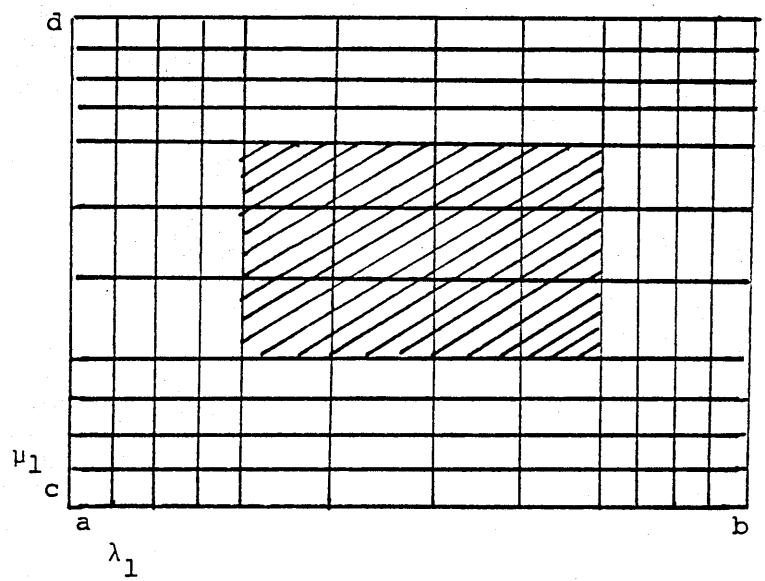


Figure 4.10

A Typical pattern of knots for a B-spline fit to the shaded region

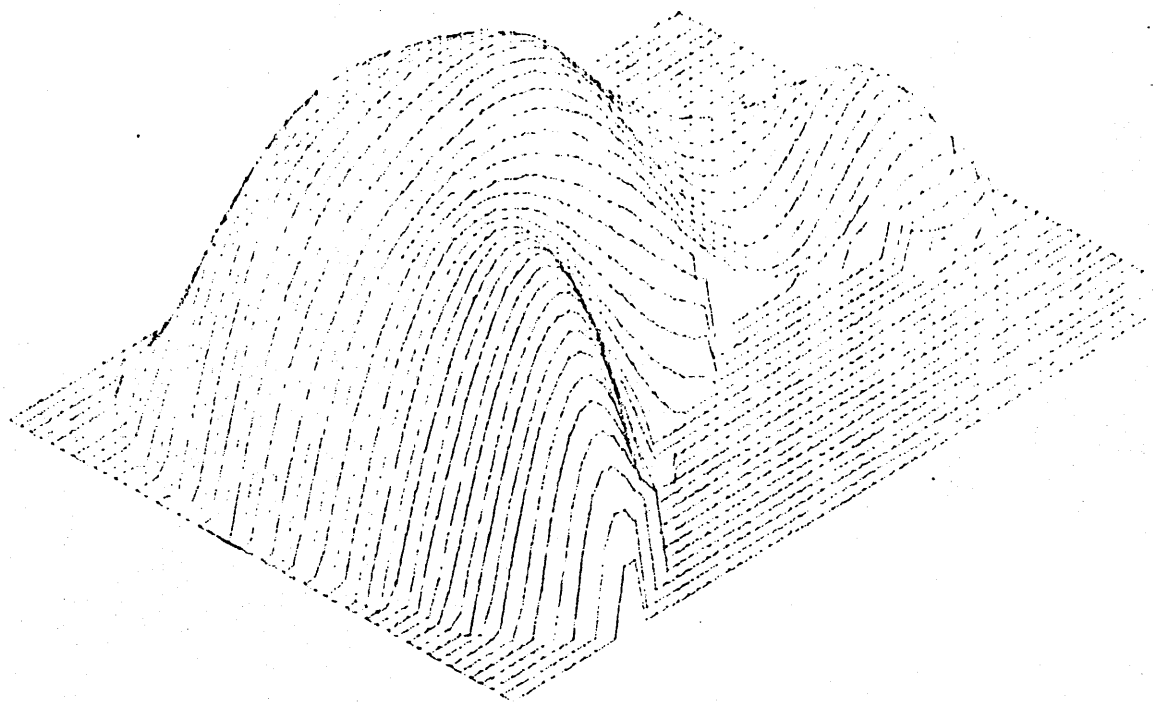


Figure 4.11

A B-spline fit to the liver scan shown in Figure 3.7

4.8 Liver Scan Characteristics

Having obtained the fitted spline surface and hence the residuals from the fit it is possible now to select the potentially important characteristics of the scans. The measures extracted from each scan were,

Property	Measure
(a) Size	number of non-negative values in the processed scan
(b) Maximum	largest fitted value
(c) Mean	average of the fitted values
(d) Total	sum of fitted values
(e) Quality	Standard deviation of the fit

As well as overall figures, the livers were divided into their constituent lobes at the point where a vertical plane through the liver has the smallest average level. The properties mentioned above were then calculated for each lobe.

The texture of the residuals should give some indication of the patchiness. Texture measurement has a long history in

image analysis and very many methods have been proposed, see Haralick, Shanmugan and Dinstein(1973) for a review. Wezka, Dyer and Rosenfeld(1976) in a comparitive study found little to choose between the various meausures when applied to terraine classification and so a comparatively simple measure based on grey scale differences has been adopted.

For fine texture the window (a) from figure 4.13 was used, the average and standard deviation over the whole liver being recorded. For coarser texture the window (b) was used. Again the mean and standard deviation over the entire liver were recorded.

To detect lesions one ideally wishes to located patches within the residuals where values are all large and negative. To facilitate this the following algorithm was employed.

Stage 1:

Smooth the residuals by use of the window,

$$\begin{matrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{matrix}$$

passing three times over the scan.

Stage 2:

Locate any patches in the resulting smooth, where a patch is defined as a collection of neighbouring points (touching along at least one side), that contain at least one value less than,

-2 x standard deviation of smooth

and all points less than

-1.5 x standard deviation of smooth

Such patches are found by first finding a point satisfying the first inequality and then searching vertically and horizontally from that point.

Whilst the smoothing of the residuals does help to define the patches it also has the effect of generating artifical patches in an otherwise random scatter of residuals. In order to distinguish the real lesions from artifacts the mean and standard deviation of the original residuals was recorded along with the size

5. FEATURE SELECTION AND EXTRACTION

5.1 Introduction

Having set the scene by introducing the data sets and describing their characterisation, we will now progress to the stage of feature selection. In chapters six and seven new methods of feature selection and extraction are proposed and tested. In order to see how these relate to the methods already in common use this chapter will review, in some detail, the most frequently used algorithms, concentrating on those relevant to a statistical view of the problem.

In his review of developments in pattern recognition Fu(1980) expressed the opinion that,

> "Strictly speaking the study of feature extraction is problem dependent"

This is perhaps an extreme view, especially if, as we have choosen to do, one differentiates between characterisation and feature selection, for there are many techniques and algorithms that have found very wide applicablity.

All scientific investigations are based on a subjective characterisation of the problem in which the investigator uses his or her own knowledge of the problem to select those

aspects of the problem that are worthy of study. Thus with ballistocardiograms it is common practice to concentrate one's attention on the peaks of the waveform. However after that stage there is scope for the use of automatic procedures for selecting out those features that best discriminate between the classes.

It is not uncommon to find that the initial data set resulting from a characterisation of the data contains tens, hundreds or even thousands of variables so that without some automatic procedure there would be little hope of sifting out the most important features. The need to reduce an initially large set of variables is usually a result of the large number of parameters that would need to be estimated if a discriminant function were to be produced using them all. Such an enormous estimation problem would itself require such large data sets that it would never be practical even assuming that the computing power were available.

Automatic feature selection methods are of two main types, either an attempt is made to select the best subset of the variables, or some reduced number of new features are created by transforming the original data set. These two approaches are reviewed separately but we start in section 5.2 by looking at the criteria that might be used for judging what is the best subset.

5.2 Feature Selection Criteria

The obvious definition of the best subset of variables in a discrimination problem is that which produces the smallest classification error. This error is itself dependent on the choice of classifier but would ideally be linked to the best possible, or Bayes classifier. Unfortunately, as mentioned in the general review in chapter two, the probability of error is an extremely difficult quantity to estimate, existing methods, in all but the simplest cases, being time consuming and relatively unreliable.

The difficulty in estimating the Bayes error has lead to a search for other criteria that are easy to evaluate but which are linked in some way to the Bayes error. Numerous functions, D(1:2), have been proposed for measuring the distance between two classes w_1 and w_2. Such distances are usually required to have propeties,

(i) $D(1:2)=0$ if $f(\underline{x} \mid w_1)=f(\underline{x} \mid w_2)$

(ii) $D(1:2) \geq 0$

(iii) $D(1:2)=D(2:1)$

(iv) $D(1:2)$ is maximised for disjoint classes.

In the two class case the Bayes probability of error, P_e, could be expressed as,

$$P_e = \tfrac{1}{2}\left|1 - \int |P(w_1|\underline{x}) - P(w_2|\underline{x})|\, f(\underline{x})\, d\underline{x}\right|$$

where $f(\underline{x})$ is the density of the mixture, that is,

$$f(\underline{x}) = P(w_1) f(\underline{x}|w_1) + P(w_2) f(\underline{x}|w_2)$$

This leads to the natural measure of distance known as the Kolmogorov variational distance.

$$D_{kv}(1:2) = \int |P(w_1|\underline{x}) - P(w_2|\underline{x})|\, f(\underline{x})\, d\underline{x}$$

However whist this distance is directly linked to the probability of error it is not simple to evaluate and one is little further on.

Table 5.1 gives a summary of some of the distance measures that have been proposed in the engineering literature for use in feature selection. They are given in their two class form but may be extended to cope with multiple classes. For example, with m classes the Bhattacharyya distance would become,

Table 5.1

Some commonly used distance measures adapted from Kanal (1974)

Name	Expression
Shannon Entropy	$E\{ - \sum_{i=1}^{m} P(w_i \mid \underline{x}) \ln P(w_i \mid \underline{x}) \}$
Bayesian Distance	$E\{ \sum_{i=1}^{m} P(w_i \mid \underline{x})^2 \}$
Bhattacharyya Distance	$E\{ [P(w_1 \mid \underline{x}) \; P(w_2 \mid \underline{x})]^{\frac{1}{2}} \}$
Chernoff Distance	$E\{ P(w_1 \mid \underline{x})^{1-S} \; P(w_2 \mid \underline{x})^{S} \}$
Generalised Kolmogorov Distance	$E\{ \lvert P(w_1 \mid \underline{x}) - P(w_2 \mid \underline{x}) \rvert^{\alpha} \}$ $0 < \alpha < \infty$

$$\sum_{j=1}^{m} - \ln(P(w_j)) \int_S f(\underline{x}|w_j) f(\underline{x})\, d\underline{x}$$

For each of the suggested measures of distance it has been shown that as the distance increases so the Bayes error decreases and in many cases bounds have been found for P_e in terms of the distances. See, for example, Lainiotis and Park(1973) and Vilmansen(1973). However these bounds are not tight and one still has the problem of evaluating the integral in any particular case. It is generally true that the engineering literature virtually ignores the problems of estimation and the literature on selection criteria is a good example of this. The papers that suggest and test selection criteria all assume known distributional forms and ignore the tremendous difficulties that would arise in trying to estimate any of these distances from a set of data.

It is generally the case that these distance measures are no better than the probability of error and as Kanal(1974) suggested when considering this problem,

> "one should try to estimate the probability of error in some direct manner."

5.3 Variable Selection

In this section we will consider the problem of choosing the best subset of size q, from a set of p potential features. As we mentioned in the previous section, one first needs some criterion by which two subsets can be compared but even when such a criterion is available problems still remain. The major difficulty is the large number of subsets that need to be compared. If there are p variables to choose from then there will be ${}_pC_q$ possible subsets of size q and this quantity quickly becomes very large. For example if p=40 and q=10 then there are nearly 850 million possible subsets and an exhaustive search through them is not a practical proposition.

A subsidiary problem in variable selection is that of choosing the value for q. Commonly the size of the selected feature set is increased until it is subjectively judged to give sufficiently good discrimination. Care needs to be taken however, to ensure that the training sets are large enough to support the number of selected features; for if they are not, one may observe the phenomenon discussed in chapter two whereby, although the performance of the classifier appears to be improving, it is in fact getting

worse.

Below we describe the currently available alternatives to exhaustive search. Of these methods only the branch and bound algorithm is optimal in the sense of guaranteeing the same result as would be obtained by exhaustive search.

Sequential Forward Selection

According to this procedure one selects the best single variable, then progressively adds one variable at a time to those already selected, picking the variable that together with those already choosen optimises the selection criterion.

The main objection to this approach is that once included a variable cannot be later rejected from the subset. The subset that results may not be the best possible even when the variables are independent, as has been demonstrated by Cover(1974).

Sequential Backward Selection

Effectively this is the reverse of sequential forward selection. The starting point is the entire set of potential features from which the least important variable is deleted at each stage, stopping when the required number remain.

This method is open to a similar objection to forward sequential selection, that is, that once a variable has been rejected from the subset in can never be re-introduced. Further since the entire data set is used at the start the method is likely to be more expensive in computer time and demands that we have training sets sufficiently large to enable the classifier and selection criterion to be estimated from the full set of variables. Conversely one does at least have an indication of the performance of the entire set of variables.

'Plus s - Take Away r'

This generalisation of the sequential procedure called 'plus s - take away r' by Kittler(1978) overcomes the basic criticisms of sequential forward and backward selection by adding at each stage the best s variables and deleting the worst r variables.

This type of procedure has long been employed in stepwise regression and discriminant analysis and is a feature of many statistical packages. However it is common to continue until the value of the criterion reaches some preset limit rather than restricting oneself to a fixed number of variables.

The problem with this modification as it stands is

that although s variables are added together they are selected on the basis of their performance when added singly. It would clearly be better when making ones choice if one looked at all subsets of size s and r computing their combined effect on the selection criterion.

Generalised Sequential Forward Selection

This generalisation of sequential forward selection considers the variables in groups of size s and finds the best subset of that size for adding to the variables already choosen. This method thereby makes allowance for the relationships between the variables in the potential subsets.

Generalised Sequential Backward Selection

This is similar to the generalisation of sequential forward selection except that the r variables to be deleted at each stage are treated as a set so as to allow for relationships between them.

Generalised 'Plus s - Take Away r'

An amalagamation of the previous two, this method

finds the best subset of size s to add to those variables already selected and the least useful subset of size r to reject. In both cases the variables are assessed together as a single subset.

Max-Min Algorithm

This method, investigated by Backer and De Shipper(1977), involves considering separately each currently unselected variable paired in turn with each of those already selected. For these combinations the minimum contribution of each potential addition is noted and the variable selected is the one for which the minimum contribution is largest.

Since the variables are only ever used in pairs it is possible to look at all combinations before starting the selection and the computer time is relatively small.

Standardised Discriminant Function Coefficients

If the potential variables are first standardised by dividing by their standard deviations, then the coefficients of any linear discriminant function may be expected to reflect the relative importance of the features. The q features to be selected are thus those with the largest coefficients. Of course one needs

sufficient data to estimate the discriminant function with all the variables.

Branch and Bound

The idea behind this algorithm is that one should find a criterion D that cannot increase if the set of selected variables is decreased. The probability of error and most distances would satisfy this requirement, if one ignores the problems of estimation.

Starting with the full set of variables, the criterion is evaluated and then the variables are removed producing a tree like structure of the type illustrated in figure 5.1.

The best solution obtained to date in figure 5.1 which involves deleting two variables, is obtained when x_2 and x_3 are removed. Since at this point D=10 there is no point in pursuing any branch for which the criterion has already fallen below 10. Thus one need not consider further deletions from the branch that starts by deleting x_4. By this means interest is restricted to those branches for which the criterion value remains above the current best.

This type of algorithm has found many applications in statistics and operational research, as Hand(1981b) has described. By not requiring that all branches be

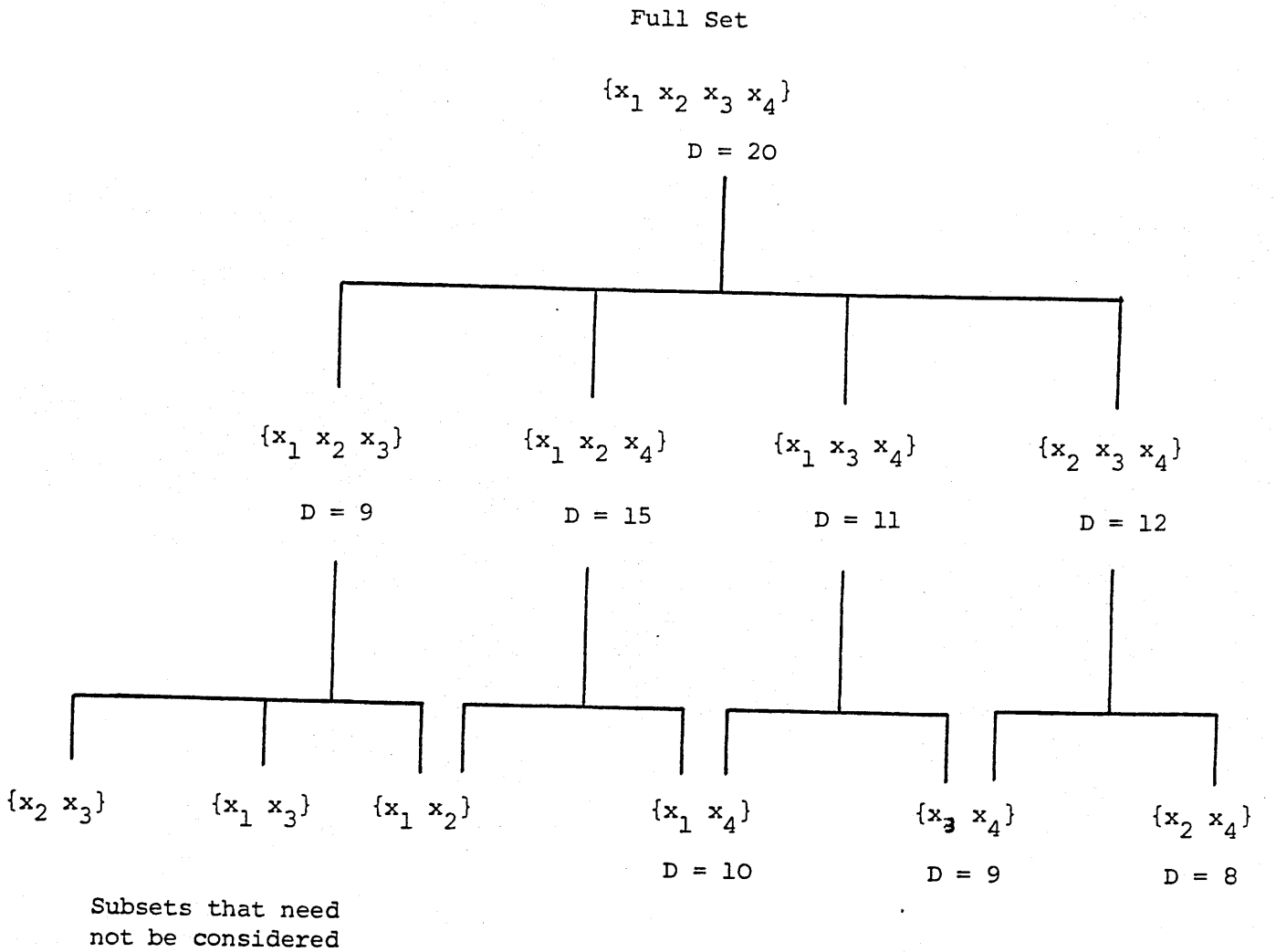


Figure 5.1

An example of the Branch and Bound Algorithm

followed the algorithm greatly reduces the number of subsets that need to be considered but at the same time guarantees to find the best subset. There are many ways in which one might decide on the order in which the branches are to be followed and obviously the one chosen will have an effect on the speed of the algorithm. Kittler(1978) gives one such method and Roberts(1984) gives a Fortran program for the branch and bound algorithm.

5.4 Stopping Rules

Rather than select the number of features, q, beforehand it would be much more sensible to allow the data to choose q. Thus when forward selection is being used one would wish to continue until adding extra features makes no significant improvement to the discrimination. Since, for most criteria, the addition of a extra variable will always make the apparent performance better, significance here must denote a improvement over and above that which one would obtain from any randomly selected and unrelated variable. Of course as with any statistical analysis significance does not denote importance and it may be necessary for the investigator to insert his own judgement into the analysis.

If one is willing to assume multivariate normality with equal covariance structures then the Mahalanobis distance is a good measure of performance and Rao(1946) gave a test for the difference between two such distances D_1^2 and D_2^2 based on q_1 and q_2 features, where the q_1 form a subset of the q_2.

The test statistic being,

$$\frac{n_1 + n_2 - q_2 - 1}{q_2 - q_1} \quad \frac{n_1\, n_2\, (D_2^2 - D_1^2)}{(n_1 + n_2)(n_1 + n_2 - 2) + n_1\, n_2\, D_1^2}$$

where n_1 and n_2 are the sample sizes. Care needs to be taken, even with this test, for as McKay(1976) points out the distributional theory derived by Rao assumes that the sets of variables are selected without reference to the data, a very unlikely assumption in practice. Thus whilst one might use this statistic as a stopping rule one should be very careful about quoting specific error rates for the test.

Costanza and Afifi(1979) compared seven stopping rules for use with forward selection. The methods tried included,

(i) testing D_q^2 against D_p^2

(ii) testing D_{q+1}^2 against D_q^2

and the minimisation of the probability of error based on different estimates of the Mahalanobis distance. As might be anticipated no method proved consistently better than the others.

5.5 Feature Extraction by Transformation

Many methods have been proposed for finding combinations of the original variables, usually linear combinations, that distinguish well between the classes. Guseman, Peters and Walker(1975) considered the general problem of finding the linear combinations that minimised the Bayes probability of error when the data came from multivariate normal distributions with unequal covariance structures. Although they did manage to derive one or two theoretical results they found that only the two class, single feature case was computationally feasible.

As with so many of the problems in discrimination the ideal solution based on minimising the probability of error, once again turns out to be prohibitively complex and alternatives must be sought. The most popular approaches rely on the principal component transformation, known in the engineering literature as the Karhunen-Loeve expansion. Principal components were first introduced around the turn of the century as a method of orthogonal regression and were later rederived by Hotelling(1930) in connection with Factor Analysis.

The basic result is that given a real symmetric positive

semi-definite p-dimensional matrix $\underline{S}$, such as a covariance matrix, then it is possible to express $\underline{S}$ in terms of its eigenvalues λ_i and normalised eigenvectors,

$$\underline{g}_i \qquad i = 1, \ldots, p$$

The expansion being,

$$\underline{S} = \sum_{i=1}^{p} \lambda_i \underline{g}_i \underline{g}_i'$$

The matrix may thus be approximated by taking terms in the expansion that correspond to the largest eigenvalues and it may be shown that, in a least squares sense, the best q-dimensional approximation to the p-dimensional data is obtained by using the space defined by the eigenvectors corresponding to the q largest eigenvalues.

Chein and Fu(1967) applied this transformation to the pooled covariance matrix and Watanabe(1965) applied it to the covariance matrix of the entire data set. The former approach will only succeed in discriminating between classes if by chance the main differences between the classes lie in the same space as the main sources of variation within classes. The latter approach is more likely to be successful since it will also work if the variation between classes is much larger than the variation within classes. Tou and Heydorn(1967) advocated the use of the eigenvectors associated with the smallest eigenvectors but this method

also only works if by chance the main discriminating information lies in that space.

Wanatabe(1965) also advocated the use of the principal components of one the classes and was thus in one sense a forerunner of the SIMCA method. This technique, described in Wald(1976), involves taking the first few principal components from each class and comparing the distances of any unclassified point to the spaces defined by those components. The point being classified into the class with the closest space.

An interesting variation on this general theme, applicable in the two class case, was suggested by Fukunaga and Koontz(1970). They first linearly transformed the variables in such a way that the pooled covariance matrix becomes the identity. They then showed that the covariance matrices of the two classes would have the same eigenvectors and that these eigenvectors would correspond to the eigenvalues given in opposite orders. Thus the eigenvector corresponding to the largest eigenvalue of one covariance matrix would correspond to the smallest eigenvalue of the other. They therefore suggested that the eigenvectors corresponding to the largest and smallest eigenvalues be chosen to define a subspace in which classification should occur. Despite the attractiveness of the method it too cannot guarantee to discriminate between classes. Foley(1973) gives an example of the method's failure.

An attempt to use the idea of principal components but also to obtain discrimination was made by Healy and Parish(1979), they first took the eigenvectors of the pooled covariance matrix to define p new variables which, of course, retain all the available information. They then formed a linear discriminant function of the Fisher type using those transformed variables. The axes with the largest coefficients were then chosen to define the space for future study. Effectively then they are trying to overcome the objections to using the largest or smallest eigenvalues by devising a method that selects the most important eigenvalues. One is not however assured of a solution that improves on the use of a selection of the original variables and whatever happens the results will be much more difficult to interpret.

A similar idea was suggested by Kittler and Young(1973) when they advocated the use of the space defined by the eigenvectors of the pooled covariance matrix standardising the new variables and then defining a covariance type matrix between the class means on the transformed scale. The new eigenvectors of this matrix are then taken as the axes for discrimination.

It should be noted that any analysis performed on the covariance matrix could also be performed on the correlation matrix. This would be equivalent to analysing the standardised data. Indeed it is a point that appears to be

frequently overlooked that the principal components are not invariant under changes of scale. Thus it would be possible to get two entirely different sets of features merely by measuring the variables in different units; A property that should not be associated with discrimination. Despite the general inappropriateness of the method it is still commonly used as a method for feature selection.

Canonical variate analysis stems directly from the original work of Fisher(1936) and also reduces to an analysis of eigenstructure. The important difference is that this analysis is specifically geared towards discrimination. Using the notation set out in appendix I, where $\underline{B}$ is the between samples sums of squares matrix and $\underline{W}$ is the within samples sums of squares matrix, then canonical variate analysis seeks to maximise,

$$\frac{\underline{u}^{\intercal}\,\underline{B}\,\underline{a}}{\underline{a}^{\intercal}\,\underline{W}\,\underline{a}}$$

This function is maximised when one takes $\underline{a}$ as being the eigenvector of $\underline{W}^{-1}\underline{B}$ corresponding to the largest eigenvalue.By choosing further eigenvecotrs, in order of the size of their eigenvalues, we obtain progressively more of the total discriminatory power.

Originally canonical variate analysis was based on the assumption of equal variation within classes so that a pooled matrix could be used for $\underline{W}$. Since then many generalisations have been propsed. Foley and Sammon(1975) looked at the maximum of,

$$\frac{\underline{a}^{'} \underline{B} \underline{a}}{\underline{a}^{'} \underline{W}^{*} \underline{a}}$$

where,

$$\underline{W}^{*} = K \underline{W}_1 + (1 - K)\underline{W}_2$$

and $\underline{W}_1$ and $\underline{W}_2$ are the two individual within sample sums of squares matrices.

K azakos(1977) investigated the maximum of

$$\frac{|\underline{a}^{'} (\underline{\mu}_1 - \underline{\mu}_2)|}{\sqrt{\underline{a}^{'} \underline{\Sigma}_1 \underline{a}} + \sqrt{\underline{a}^{'} \underline{\Sigma}_2 \underline{a}}}$$

and Kittler(1977) considered,

$$\frac{\sum_{j=1}^{m} P(w_j) (\underline{a}^{'} \underline{\mu}_j)^2}{a^{'} \underline{\Sigma} \underline{a}}$$

Fehlauer and Eisenstein(1978) suggested a clustering

algorithm based on the maximisation of,

$$\frac{\underline{a}' (\underline{B} + \underline{W}_2) \underline{a}}{\underline{a}' \underline{W}_1 \underline{a}}$$

and Gilsema and Eden(1980) generalised this to,

$$\frac{\underline{a}' (\underline{B} + \beta \underline{W}_2) \underline{a}}{\underline{a}' (\underline{W}_1 + (1-\beta) \underline{W}_2) \underline{a}} \qquad 0 \leq \beta \leq 1$$

$$\frac{\underline{a}' (\underline{B} + \beta \underline{W}_2) \underline{a}}{\underline{a}' \underline{W}_1 \underline{a}} \qquad \beta > 1$$

and described its role within an interactive package.

Along similar lines Wilks(1962) advocated the minimisation of,

$$\frac{|\underline{a}' \underline{W} \underline{a}|}{|\underline{a}' \underline{T} \underline{a}|}$$

where $\underline{T}$ is the total sums of squares matrix. His results have since been rederived in the engineering literature.

All of the eigenstructure methods make the implicit assumption of linearity. However, if quadratic effects are suspected it is perfectly possible to create new variables consisting of the squares or crossproducts of the original variables and to include them in the analysis.

One promising approach to linear feature selection that has received little attention is the projection pursuit algorithm of Friedman and Tukey(1974). The object of their method is to find subspaces, defined by linear combinations of the original variables, that maximise a score of 'usefulness', defined as the product of a measure of overall spread and a measure of local density. The algorithm thus seeks out projections in which points cluster into clearly defined and well separated groups.

Many non-linear transformations have been proposed which it is hoped would give better class separation, all be it at the expense of greatly increased computation. One of the earliest of these was a type of non-metric scaling proposed by Sammon(1969), who sought to minimise a 'stress' function dependent upon,

$$\Sigma(\delta_i - d_i)^2$$

where δ_i are the distances between objects in the original space and d_i are the distances in the reduced space. Such stress functions are now in common use in statistics and programs for their iterative minimisation are widely available. Many variations on the form have the stress function have been propsed and details may be found in any book on multidimensional scaling.

One other method of non-linear analysis deserves mention

and that is a technique propsed by Olsen and Fukunaga(1973). According to this method samples are divided into clusters of similar cases and each cluster is summarised by a well fitting subspace. These sub-spaces are then combined to form a space describing the whole sample.

5.6 Comparisons of Feature Selection Techniques

With such a large number of feature selection techniques available there is clearly a need for some guidelines on the choice of method for a particluar problem and yet virtually none exist. The difficulty is that it is very hard to compare methods on real data because of the other factors that come into play. Most important of these being the choice of classifier and the assessment of the schemes performance. Even when such a comparison is made one has no assurance that the results will carry across to other data sets. Thus one is left with only one option namely to look at simulated, usually multivariate normal data. Again the problem is that little is known about the robustness of normal based feature selection methods.

Mucciardi and Gose(1971) gave one example of a comparative study based on ECG data. They compared a variety of approaches including forward selection based on the univariate probability of error and forward selection based on the average correlation between a potential variable and those already selected. It is not surprising that their conclusions are vague. All methods were found to be better

than random selection of the features and reliance on univariate charcteristics was found to give poor results.

Weiner and Dunn(1966) also conducted a small scale study including random selection of the features as a type of control. No method proved consistently better than any other and no firm conclusions were drawn.

Kittler(1978) performed a more useful study of variable selection methods using two 20-dimensional normal populations with identical covariance structures. He was thus able to use the Mahalanobis distance between classes as a genuine measure of absolute performance. His conclusions were that the generalised methods described in section 5.3 perform better than simple forward or backward selection and that the max-min algorithm performs poorly.

6 A NEW METHOD OF FEATURE SELECTION BASED ON THE ELIMINATION OF CASES

6.1 Introduction

Having reviewed the currently available methods for feature selection we will now consider a new method suitable for large scale applications. The proposed method is very simple to program and flexible enough to be adapted for use with most types of data.

The procedure is sequential in the manner of those methods described in section 5.3, but with the advantage that one only ever needs to consider univariate distributions. This desirable simplification is achieved by conditioning the selection on unhelpful values for the previously chosen variables. When nothing is known about the forms of the distributions this can be achieved by, at each stage, eliminating those cases that would be clearly classified by the chosen variable used singly. As a result of this elimination the training sets will get progressively smaller until eventually they are completely used up; at that point the feature selection must stop.

The suggested method will be shown to have the advantages resulting from the use of univariate criteria but to out perform selection based solely on the marginal distributions.

This method is, like any other, dependent upon the choice of a suitable selection criterion, but it also requires a decision rule for deciding on those cases that would be clearly classified.

In order to look at the performance of the scheme under something like standard conditions we will investigate its use with multivariate normal data. However the primary strength of the method lies in its flexibility and it may equally well be applied to non-normal data or with a non-parametric criterion. In the final section of the chapter the method is applied to the data on ballistocardiograms and the results are compared with those obtained in an earlier, more conventional, analysis.

6.2 Conditional Selection with Normal Data

Suppose that we are trying to discriminate between two p-dimensional normal distributions, $N(\underline{\mu}_1, \underline{\Sigma})$ and $N(\underline{\mu}_2, \underline{\Sigma})$ and that we have already, for whatever reason, chosen the first q variables for inclusion in the analysis. We now wish to select one more variable for inclusion in the analysis.

Clearly the ideal solution to this problem is based on the Mahalanobis distances between the distributions using the q selected variables and each of the other potential additions. The variable that, when used in addition to the original q variables, gives the largest separation being the one that would be chosen.

If one wanted to avoid the multivariate structure of this problem then one might use the marginal distributions of each potential addition. In this case the Mahalanobis distance would reduce to a type of statistic that will be called t^2

$$\frac{(\mu_{1i} - \mu_{2i})}{\sigma_i^2} = \frac{d_i^2}{\sigma_i^2}$$

From the p-q possible additions the one with the largest value of t^2 would be the one chosen. Obviously such a

selection would take no account of the correlations between the q variables already selected and the potential additions; consequently the selection would often be a poor one.

Suppose however that one were to look at the distributions of the potential additions conditional upon the previously selected variables taking some unhelpful value $\underline{b}$,

$$\text{i.e. } f(x_i \mid \underline{x}_q = \underline{b})$$

These distributions will be univariate normal with means,

$$\mu_{1i} + \underline{r}_{qi}' \underline{\Sigma}_q^{-1} (\underline{b} - \underline{\mu}_1)$$

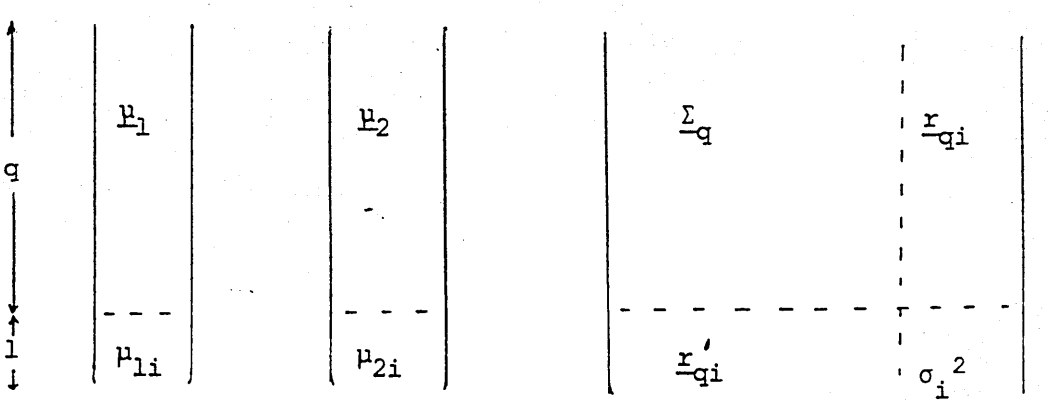
and

$$\mu_{2i} + \underline{r}_{qi}' \underline{\Sigma}_q^{-1} (\underline{b} - \underline{\mu}_2)$$

and common variance,

$$\sigma_i^2 - \underline{r}_{qi}' \underline{\Sigma}_q^{-1} \underline{r}_{qi}$$

where the means and variances of the set of q+1 variables have been partitioned so that,



Now the conditional t^2-statistic between the two distributions is,

$$t^2 = \frac{(d_i - \underline{r}_{qi}' \underline{\Sigma}_q^{-1} \underline{d})^2}{\sigma_i^2 - \underline{r}_{qi}' \underline{\Sigma}_q^{-1} \underline{r}_{qi}}$$

where,

$$d_i = \mu_{i1} - \mu_{i2}$$

and

$$\underline{d} = \underline{\mu}_1 - \underline{\mu}_2$$

In this rather special case our criterion does not depend on the conditioning value b, this will generally not be the case.

The benefit to be derived from the use of the conditioned distributions is most easily seen by considering the first few stages in the selection process. Suppose that q=1 and that we are now seeking a second variable. To simplify the notation assume, without any loss of generality, that the previously selected variable is x_1 and that all of the variables have been scaled to have unit variance.

The Mahalanobis distance based on variables x_1 and x_i is thus,

$$D_{1i}^2 = \underline{d}' \underline{\Sigma}^{-1} \underline{d}$$

where,

$$\underline{d} = \begin{bmatrix} d_1 \\ \\ d_i \end{bmatrix} \qquad \underline{\Sigma} = \begin{bmatrix} 1 & \rho_{1i} \\ \\ \rho_{1i} & 1 \end{bmatrix}$$

Clearly,

$$\overline{D^2_{1i}} = \frac{d_1^2 - 2\rho_{1i}\, d_1\, d_i - d_i^2}{1 - \rho_{1i}^2}$$

and the conditional t^2-statistic is,

$$t^2 = \frac{(d_i - \rho_{1i}\, d_1)^2}{1 - \rho_{1i}^2} = D^2_{1i} - d_1^2$$

Thus our conditional t-statistic is equal to the increase in the Mahalanobis distance and will thus give optimal results.

Using the natural extensions of this notation we might obtain similar expressions in the case where $\alpha=2$. Here supposing that we have already selected x_1 and x_2 then,

$$\begin{aligned} D^2_{12i} = \{&(1 - \rho_{2i}^2)\, d_1^2 + (1 - \rho_{1i}^2)\, d_2^2 + (1 - \rho_{12}^2)\, d_i^2 \\ &- 2(\rho_{12} - \rho_{1i}\rho_{2i})\, d_1 d_2 - 2(\rho_{1i} - \rho_{12}\rho_{2i})\, d_1 d_i \\ &- 2(\rho_{2i} - \rho_{12}\rho_{1i})\, d_2 d_i \} \\ &/ (1 - \rho_{12}^2 - \rho_{1i}^2 - \rho_{2i}^2 + 2\rho_{12}\rho_{1i}\rho_{2i}) \end{aligned}$$

and

$$t^2 = \{d_i(1 - \rho_{12}^2) - \rho_{1i}\,(d_1 - \rho_{12}\,d_2) - \rho_{2i}\,(d_2 - \rho_{1i}\,d_1)\}^2$$

$$/\ (1 - \rho_{12}^2 - \rho_{1i}^2 - \rho_{2i}^2 + 2\rho_{12}\,\rho_{1i}\,\rho_{2i})\ (1 - \rho_{12}^2)$$

For these two criteria to give the same solution in all cases it would be necessary to find that all terms dependent on i should have the same value. Table 6.1 compares such terms, ignoring the common divisor.

It will be seen that the two methods do not now give the same results although, for many applications, it would be reasonable to assume that,

$$d_i > \rho_{1i}\,\rho_{2i}$$

so that the two criteria will agree in the dominant terms, and hence will often give the same selection.

These results suggest what proves to be a general result, namely that if one ignores terms containing the squares or products of correlations involving the variable under consideration, then the conditional t^2-statistic is equivalent to the Mahalanobis distance as far as order of selection is concerned.

Table 6.1

A comparison of the coefficients of D^2 and t^2

TERM	COEFFICIENT	
	D^2	t^2
d_i^2	$(1 - \rho_{12}^2)$	$(1 - \rho_{12}^2)$
$\rho_{1i}\ d_i$	$-2(d_1 - \rho_{12}\ d_2)$	$-2(d_1 - \rho_{12}\ d_2)$
$\rho_{2i}\ d_i$	$-2(d_2 - \rho_{12}\ d_1)$	$-2(d_2 - \rho_{12}\ d_1)$
$\rho_{1i}\ \rho_{2i}$	$2d_1\ d_2$	$\frac{2(d_1 - \rho_{12}\ d_2)(d_2 - \rho_{12}\ d_1)}{(1 - \rho_{12}^2)}$
ρ_{1i}^2	$-d_2^2$	$\frac{(d_1 - \rho_{12}\ d_2)^2}{(1 - \rho_{12}^2)}$
ρ_{2i}^2	$-d_1^2$	$\frac{(d_2 - \rho_{12}\ d_1)^2}{(1 - \rho_{12}^2)}$

THEOREM:

Ignoring squares and product of correlations and using the already established notation,

$$t^2 = D^2_{(q)i} - D^2_{(q)}$$

Proof:

If q variables have already been selected then the conditional t^2-statistic associated with variable x_i is,

$$t^2 = \frac{(d_i - \underline{r}'_{qi} \underline{\Sigma}_q^{-1} \underline{d})^2}{1 - \underline{r}'_{qi} \underline{\Sigma}_q^{-1} \underline{r}_{qi}}$$

which, ignoring sums and products of correlations involving variable x_i, becomes,

$$d_i^2 - 2 d_i \underline{r}'_{qi} \underline{d}$$

On the other hand,

$$D^2_{(q)i} = (\underline{d}' \; d_i) \begin{bmatrix} \underline{\Sigma}_q & \underline{r}_{qi} \\ \underline{r}_{qi}' & 1 \end{bmatrix}^{-1} \begin{bmatrix} \underline{d} \\ d_i \end{bmatrix}$$

Now

$$\begin{bmatrix} \underline{\Sigma}_q & \underline{r}_{qi} \\ \underline{r}_{qi}' & 1 \end{bmatrix}^{-1} = \begin{bmatrix} \underline{A}^{-1} & -\underline{A}^{-1} \underline{r}_{qi} \\ -\underline{r}_{qi}' \underline{A}^{-1} & 1 - \underline{r}_{qi}' \underline{A}^{-1} \underline{r}_{qi} \end{bmatrix}$$

where

$$\underline{A} = \underline{\Sigma}_q - \underline{r}_{qi} \underline{r}_{qi}'$$

so that ignoring squares and products of correlations

$$\begin{bmatrix} \underline{\Sigma}_q & \underline{r}_{qi} \\ \underline{r}_{qi}' & 1 \end{bmatrix}^{-1} = \begin{bmatrix} \underline{\Sigma}_q^{-1} & -\underline{\Sigma}_q^{-1} \underline{r}_{qi} \\ -\underline{r}_{qi} \underline{\Sigma}_q^{-1} & 1 \end{bmatrix}$$

and consequently

$$D^2_{(q)i} = \underline{d}' \underline{\Sigma}_q^{-1} \underline{d} - 2 \, d_i \, \underline{r}_{qi}' \underline{\Sigma}_q^{-1} \underline{d} + d_i^2$$

so that

$$t^2 = D^2_{(q)i} - \underline{d}' \underline{\Sigma}_q^{-1} \underline{d}$$

$$= D^2_{(q)i} - D^2_{(q)}$$

It will be seen that the Mahalanobis distance criterion, the conditional t^2-statistic and the marginal t^2-statistic form a progression. Mahalanobis distance is optimal and takes full account of the correlations between variables. It is equivalent to the conditional t^2-statistic if one ignores squares and products of correlations and it is equivalent to the marginal t^2-statistic if one ignores all correlations.

There are too many factors that might affect conditional selection for one to say more than that the smaller the correlations between variables the better that the conditional selection will perform. However, in order to get some feel for the performance of conditional selection we will consider a simple example.

Suppose that, using the already established notation one has two normal distributions with,

$$\underline{d}' = (1.5, 1.5)$$

$$\underline{\Sigma}_2 = \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix}$$

$$d_i = 1.2 \qquad d_j = 1.0$$

$$\underline{r}_{2i} = (r_{1i}, r_{2i})$$

$$\underline{r}_{2j} = (r_{1j}, r_{2j})$$

The marginal criterion always selects variable i, but the optimal selection will depend upon the correlations r. In order to investigate this dependence 500 simulations were performed selecting the correlations as uniform(-1,1) variables and rejecting cases for which the corresponding covariance matrix was not positive definite. Table 6.2 shows the results of the simulation.

Table 6.2

Results of the Simulation

		t^2 choice		
		x_i	x_j	
Best	x_i	227	33	260
	x_j	28	212	240
		255	245	500

The conditional criterion makes the correct choice in 88% of trials, whereas the marginal selection, which always chose variable i, was correct only 52% of the time, little better than pure guess work. Perhaps more important is the fact that conditional selection tends to make mistakes only when the discrimination obtainable from the two possible variables is very similar and not when there is a clear difference. This is exactly what one wants in practice, for it doesn't matter so much if one chooses wrongly between an evenly matched pair so long as one doesn't miss out on the really good classifier.

Running the same simulation with a larger difference between variables i and j, that is to say with $d_j=1.2$ and

d_j=0.8, was found to make very little difference to the results as tables 6.3 shows.

Table 6.3

Results of the second simulation

	t^2 choice		
	x_i	x_j	
Best x_i	261	27	288
x_j	45	167	212
	306	194	500

The increasing difference in the marginal difference causes the conditional selection to opt for variable i rather more often than it should, but it is still achieving 86% correct selection.

It is interesting to note the similarity between this type of selection and the corresponding procedure in regression. As Anderson(1958) shows, our two class discrimination problem can be formulated as a regression of x on y where,

$$y_i = n_2/(n_1 + n_2) \text{ if from class } w_1$$

$$y_i = -n_1/(n_1 + n_2) \text{ if from class } w_2$$

Our method of conditional selection can thus be seen to be equivalent to using the partial correlation between y and x_i given the already selected variables $x_1, \ldots, x_q$ instead of the multiple correlation between y and $x_1, \ldots, x_q, x_i$. Whilst our policy is not the best possible, common sense suggests that it should perform well in most cases.

6.3 Extension to Other Known Distributions

The main problem when extending this method to other specified distibutions lies in the difficulty of defining a suitable selection criterion to act as a standard against which the conditioning method can be judged. One would like to base one's notion of the optimal selection on the probability of error but, for multivariate distributions, it is almost prohibitively expensive to calculate. Unless one is willing to pay the price of this calculation the conditioning procedure can only be compared with other sub-optimal procdures.

To see how the method would work suppose that we have two multivariate normal distributions $N(\underline{\mu}_a, \underline{\Sigma}_a)$ and $N(\underline{\mu}_b, \underline{\Sigma}_b)$ and that we have already selected variable x_1 for inclusion in a discriminant analysis and now we seek a second variable. The method requires,

(a) a univariate selection procedure,

(b) a definition of an unhelpful value.

Since we only ever need to deal with univariate distributions it would be quite possible to use the probability of error as the criterion and to define the

least helpful value as that point with equal likelihood under the two distributions that lies between the means. However, other computationally simpler choices are possible. For instance, we might choose a criterion of the form,

$$\frac{\text{difference in means}}{\text{average standard deviation}}$$

and a servicable 'unhelpful' value would be,

$$\frac{\sigma_{al}\,\mu_{bl} + \sigma_{bl}\,\mu_{al}}{\sigma_{al} + \sigma_{bl}}$$

A little algebra shows that the conditional distributions of x_i would then be normal with means,

$$\mu_{ai} + \rho_{ali}\,\sigma_{ai}\,(\mu_{bl} - \mu_{al})$$

$$\mu_{bi} + \rho_{bli}\,\sigma_{bi}\,(\mu_{al} - \mu_{bl})$$

and variances,

$$\sigma_{ai}^2\,(1 - \rho_{ali}^2) \qquad \sigma_{bi}^2\,(1 - \rho_{bli}^2)$$

so that the criterion of selection becomes,

$$4\left[\frac{d_i - d_1(\rho_{ali}\,\sigma_{ai} + \rho_{bli}\,\sigma_{bi})}{\sigma_{ai}\sqrt{1-\rho_{ai}^2} + \sigma_{bi}\sqrt{1-\rho_{bli}^2}}\right]^2$$

If x_2 were the variable selected at this stage then the next variable would be chosen conditional on,

$$x_1 = \frac{\sigma_{a1}\,\mu_{b1} + \sigma_{b1}\,\mu_{a1}}{\sigma_{a1} + \sigma_{b1}}$$

$$x_2 = \{\sigma_{a1}\sqrt{1-\rho_{a12}^2}\,[\mu_{b2} + \rho_{b12}\,\sigma_{b1}\,(\mu_{a1} - \mu_{b1})]$$

$$+ \sigma_{b1}\sqrt{1-\rho_{b12}^2}\,[\mu_{a2} + \rho_{a12}\,\sigma_{a1}\,(\mu_{b1} - \mu_{a1})]$$

$$/\{\sigma_{a1}\sqrt{1-\rho_{a12}^2} + \sigma_{b1}\sqrt{1-\rho_{b12}^2}\}$$

In this way selection would continue until sufficient variables had been selected.

It is not possible to anticipate all of the models that might be used but the general approach is clear; if one has a model for the distribution of the data then one may either use the probability of error criterion together with the point of equal likelihood , or define some measure of separation geared to that specific distribution.

6.4 Non-Parametric Selection

It is perhaps the most important aspect of this approach that it is able to encompass non-parametric selection, for it is still the case that many pattern recognition problems are simply too large to enable a study of the distributional properties of the data. If the forms of the density are completely unknown then one might estimate them non-parametrically. Then probability of error could then be used as the selection criterion and we could essentially progress as in section 6.3. Alternatively one might define some robust measure of separation and use that. Whatever choice is made this method has the great advantage that it only ever deals with univariate distributions and so the computations are kept to a minimum.

The difficulty is of course at the stage of conditioning. If we were to progress as before conditioning on $x_1=b$ then since we know nothing about the multivariate structure we must estimate the conditional distribution using cases from the training sets. Equally clearly we will not, in practice, have sufficient data for which $x_1=b$ and we will need to modify the conditioning stage so that,

$$b_{1l} < x_1 < b_{1u}$$

The values b_{1l} and b_{1u} will then be a compromise between the need for sufficient data with which to estimate the conditional densities and the desire to condition only on unhelpful values.

Our non-parametric method can thus be seen as a generalisation of sequential forward selection. After selecting the criterion and limits for the conditioning, the steps are,

(a) Calculate the value of the selection criterion for each variable and choose the best.

(b) Locate and eliminate all cases outside the conditioning interval.

(c) If the training sets are empty or if sufficient variables have been selected then stop, else one should return to (a) using the reduced training sets.

Whilst this method has been developed with two populations this is not an essential restriction. Providing that one has a univariate selection criterion measuring the separation between all classes together with a definition of unhelpful values, possibly over several disjoint intervals, then the algorithm could be applied without modification.

Because the training sets are reduced in size at each stage there will come a point when insufficient data remain for meaningful selection to continue. This might at first sight seem like a disadvantage but it is undoubtably true that, in many applications, far more variables are selected

that can be justified given the sizes of the training sets. If the elimination algorithm is used then this is guarded against in a very natural way. Finally it should be noted that the selection process will become less reliable as the training sets reduce in size and it would be advisable to stop well before the sets become completely exhausted.

6.5 The effect of using a conditioning range

Before considering the use of the elimination algorithm it is important to ensure that the use of a conditioning range will not destroy all the useful properties of this method of feature selection. We have seen that the use of conditional distributions is very effective when the two classes are normally distributed and have equal covariance structures. In this section we return to that basic set up and ask the question whether or not the use of a conditioning range instead of a single value would effect the performance of the feature selection scheme.

Our problem is then, as follows. Suppose that we have two multivariate normal classes and that to date we have selected q variables $\underline{x}_{(q)}$ and that we now seek a further variable to add to the set. We partion the means and covariances of the distributions to show the terms already selected and a potential addition x_i. Thus,

variables $\left|\begin{array}{c} \underline{x}_{(q)} \\ \hline x_i \end{array}\right|$ means $\left|\begin{array}{c} \underline{0} \\ \hline 0 \end{array}\right|$ $\left|\begin{array}{c} \underline{\mu}_{(q)} \\ \hline \mu_i \end{array}\right|$

Covariance matrix

$$\left|\begin{array}{cc} \underline{\Sigma}_{(q)} & \underline{\gamma}_{(q)i} \\ \underline{\gamma}'_{(q)i} & \sigma_i^2 \end{array}\right|$$

Further suppose that the previous stages in the selection process have lead us to condition upon,

$$\underline{L}_1 < \underline{x}_{(q)} < \underline{L}_2$$

Now the conditional distributions of the variable x_i in the two classes w_j, $j=1,2$, may be derived as,

$$f_j(x_i \mid \underline{L}_1 < \underline{x}_{(q)} < L_2) = \frac{\int_{\underline{L}_1}^{\underline{L}_2} f_j(x_i, \underline{x}_{(q)})\, d\underline{x}}{\int_{\underline{L}_1}^{\underline{L}_2} f_j(\underline{x})\, d\underline{x}}$$

Using $\phi_q(\underline{\mu},\underline{\Sigma})$ to stand for the q-dimensional multivariate normal with parameters $\underline{\mu}$ and $\underline{\Sigma}$ and letting,

$$\Phi_q(\underline{L};\underline{\mu},\underline{\Sigma}) = \int_{-\infty}^{L_q} \cdots \int_{-\infty}^{L_1} \phi_q(\underline{\mu},\underline{\Sigma})\, dx_1 \cdots dx_q$$

Then we have that,

$$f_1(\underline{x}) = \phi_q\ (\underline{0},\underline{\Sigma})$$

$$f_2(\underline{x}) = \phi_q\ (\underline{\mu},\underline{\Sigma})$$

So that

$$\int_{\underline{L}_1}^{\underline{L}_2} f_1(\underline{x})\ d\underline{x} = \Phi_q(\underline{L}_2;\underline{0},\underline{\Sigma}) - \Phi_q\ (\underline{L}_1;\underline{0},\underline{\Sigma})$$

and

$$\int_{\underline{L}_1}^{\underline{L}_2} f_2(\underline{x})\ d\underline{x} = \Phi_q(\underline{L}_2;\underline{\mu},\underline{\Sigma}) - \Phi_q\ (\underline{L}_1;\underline{\mu},\underline{\Sigma})$$

Also, by factorising we have,

$$f_1(x_i,\underline{x}_{(q)}) = \phi_1\ (0,\sigma_i^2)\ \phi_q\ (\underline{\gamma}_{(q)i}\ x_i/\sigma_i\ ,\ \frac{\underline{\Sigma}_{(q)} - \underline{\gamma}_{(q)i}\ \gamma_{q(i)}}{\sigma_i^2}$$

and

$$f_2(x_i,\underline{x}_{(q)}) = \phi_1(\mu_i,\sigma_i^2)\ \phi_q\ (\underline{\mu}_{(q)} + \underline{\gamma}_{(q)i}\ \frac{(x_i - \mu_i)}{\sigma_i}\ \frac{\underline{\Sigma}_{(q)} - \underline{\gamma}_{(q)i}\ \gamma_{(q)}}{\sigma_i^2}$$

so that

$$\int_{\underline{L}_1}^{\underline{L}_2} f_1(x_i,\underline{x}_{(q)})\ d\underline{x}_{(q)} = \phi_1\ (0,\sigma_i^2)\ \{$$

$$\Phi(\underline{L}_2;\underline{\gamma}_{(q)i}\ \frac{x_i}{\sigma_i}\ ,\ \underline{\Sigma}_{(q)} - \frac{\underline{\gamma}_{(q)i}\ \underline{\gamma}'_{(q)i}}{\sigma_i^2}$$

$$- \Phi(\underline{L}_1;\underline{\gamma}_{q(i)}\ \frac{x_i}{\sigma_i}\ ,\ \underline{\Sigma}_{(q)} - \frac{\underline{\gamma}_{(q)i}\ \underline{\gamma}'_{(q)i}}{\sigma_i^2}\ )\}$$

and

$$\int_{\underline{L}_1}^{\underline{L}_2} f_2(x_i, \underline{x}_{(q)})\, d\underline{x}_{(q)} = \phi_1(\mu_i, \sigma_i^2)\{$$

$$\Phi_q(\underline{L}_2; \underline{\mu}_{(q)} + \underline{\gamma}_{(q)i} \frac{(x_i - \mu_i)}{\sigma_i}, \underline{\Sigma}_{(q)} - \frac{\underline{\gamma}_{(q)i}\, \underline{\gamma}'_{(q)i}}{\sigma_i^2})$$

$$- \Phi_q(\underline{L}_1; \underline{\mu}_{(q)} + \underline{\gamma}_{qi} \frac{(x_i - \mu_i)}{\sigma_i}, \underline{\Sigma}_{(q)} - \frac{\underline{\gamma}_{(q)i}\, \underline{\gamma}'_{(q)i}}{\sigma_i^2})\}$$

In order to illustrate the effect of conditioning upon a range we return to the set up used in section 6.2 which when showed then benefits of conditioning upon a single unhelpful value. Here we have x_1 and x_2 as two already selected variables and x_i as the potential addition. We suppose that the means and variances are,

$$\begin{pmatrix} x_1 \\ x_2 \\ x_i \end{pmatrix} \quad \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \quad \begin{pmatrix} 1.5 \\ 1.5 \\ 1.2 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 & .5 \\ 0 & 1 & .7 \\ .5 & .7 & 1 \end{pmatrix}$$

and that the conditioning of

x_1 and x_2 is based upon,

$$-.5 < x_1 < 2.$$

$$-.5 < x_2 < 2.$$

These figures are purely illustrative but not unrealistic. The forms of the conditional distributions are shown in figure 6.1 and it will be noted that they are not markedly non-normal. This is verified by an inspection of the properties based on their first four moments as shown in table 6.4.

Table 6.4

Properties of the conditional distributions

mean	.078	.591
st deviation	.871	.784
skewness	.038	-.242
kurtosis	-.094	.024

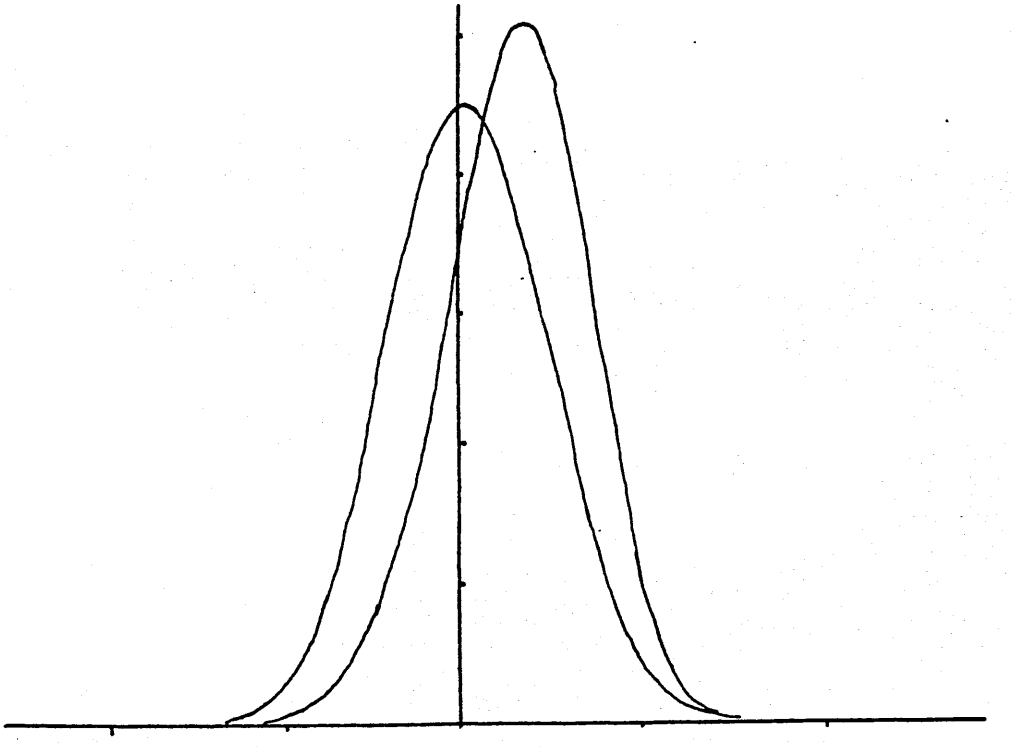


Figure 6.1

Two distributions conditioned on a range of values

In order to gauge the effect of the elimination algorithm the simulation described in section 6.2 was repeated, only this time conditioning on a range of values. Since we are dealing with univariate distributions with known, if awkward, forms there is no problem in numerically evaluating the probability of error for each potential addition. However the distibution shown in figure 6.1 suggests that we would have got very similar results from using a t^2 statistic.

In the first re-run of the simulation the conditioning ranges were set at,

$$-.5 < x_1 < 2.$$

$$-.5 < x_2 < 2.$$

and as before two potential variables x_i and x_j were compared. The first, x_i had a marginal separation of 1.2 and the second a marginal separation of 1.0, thus any purely marginal scheme would always choose x_i. The correlations were as before randomly generated and the Mahalanobis distance used as a yardstick by which to assess perfect choice.

Table 6.5 shows the results of this simulation and it will be seen that the use of a range has had very little adverse effect. Previously the method had selected the inferior variable on 12% of occassions whereas it is now in error on 13.6% of trials. The loss due to using the range is

thus very small.

Table 6.5

Simulation using a conditioning range

		Conditional x_i	Conditional x_j	
Best	x_i	239	20	259
Selection	x_j	68	173	241
		307	193	500

Changing the widths of the intervals slightly has very little effect on the results. Table 6.6 shows the results when the intervals used are,

$$0. < x_1 < 1.5$$

$$0. < x_2 < 1.5$$

and the difference is negligible.

Table 6.6

Simulation with a conditioning range

		Conditional x_i	Conditional x_j	
Best	x_i	229	35	264
Selection	x_j	56	180	236
		285	215	500

However, in the extreme if one widens the interval too far it is almost like no conditioning at all. Thus table 6.7 shows the results when the ranges are,

$$-2. < x_1 < 3.5$$

$$-2. < x_2 < 3.5$$

Now the scheme acts just like marginal selection always choosing x_i. Clearly we must be careful not to widen the intervals too far.

Table 6.7

Simulation with wide intervals

		Conditional		
		x_i	x_j	
Best	x_i	265	0	265
Selection	x_j	235	0	235
		500	0	500

6.6 Application to the Ballistocardiograms

Treated as a raw data set each ballistocardiogram contains values at 100 points. It is our object to identify those points along the wave that best distinguish between normal and pathological cases. The simplest discriminatory functions are linear and these work best for classes that differ in their means rather than their variances. An inspection of the ballistocardiogram data however suggests that the major difference between the classes of normals and abnormals is in the variability of the waves within each class. Thus figure 6.2 shows the range of amplitudes observed at the 61st point of the wave.

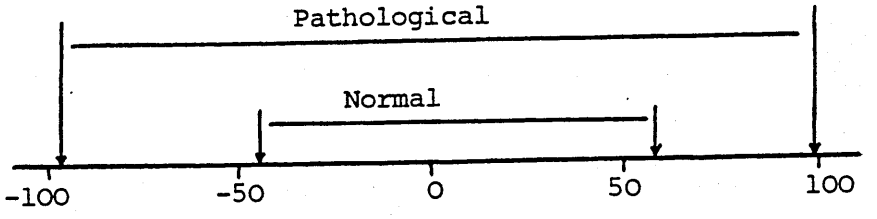


Figure 6.2

The range of amplitudes found at the 61st position in the normal and pathological groups

In order to accomodate such data one has two options either to use quadratic discriminatory functions or to transform the data before it is used. The second of these options is the simplest and was used here. Before the data were analysed they were transformed and for each variable the measurements were replaced by their absolute distances from the mid-range of the normal class. The result of this is that pathological ballistocardiograms are now characterised by larger values.

The distributions of our transformed variables are by no means normal and we therefore require some non-parametric method of analysis. Density estimation could have been used but this is time consuming and open to question over the choice of smoothing parameter and so a simple univariate non-parametric measure of separation was used.

The measure of separation was defined as,

$$R = \frac{\min(\text{pathological}) - \max(\text{normal})}{\max(\text{pathological}) - \min(\text{normal})}$$

This is essentially the proportion of the overlap between the two classes and further details of its properties may be found in chapter nine.

The elimination rule used to condition the selection was to remove from the training sets any case that would be correctly classified by that variable used singly. Thus the

second variable is selected conditional upon,

Min(pathological) < x < Max(normal)

The whole algorithm is very simply programmed. One takes each of the 100 variables in turn and calculates the extremes of the two training sets. These extremes define both the measure of separation and the elimination rule. The variable with the largest value of R is selected and any case that would be correctly classified by that variable is removed from the training sets. The extremes of the conditional data are then found and used in the next stage. The process being repeated until no cases remained.

Application of this method led to the selection of five variables, these were in the order of selection,

Set (i) 97 68 30 56 62

The positions of these variables on a normal ballistocardiogram can be seen from figure 6.3 and it will be seen that we have picked out,

(a) the peak of the J wave

(b) the beginning, peak and end of the MNO sequence

(c) the degree of damping at the end of the cycle.

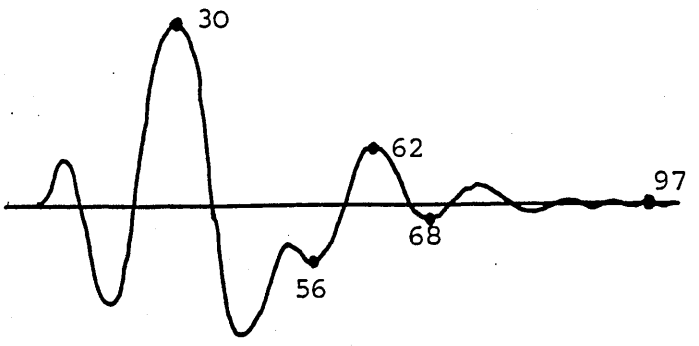


Figure 6.3

Selected features shown on a normal Bcg

If one chooses to widen the conditioning interval so that relatively more data sre retained then the selection changes; but not completely. The intervals were widened by 5% at each side as shown in figure 6.4 and then method re-run.

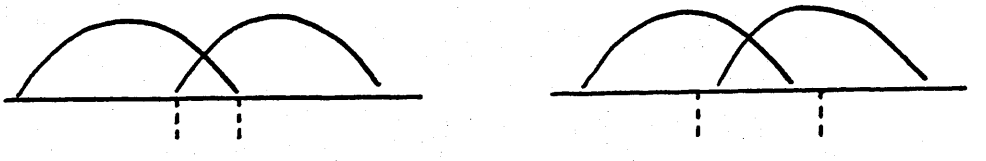


Figure 6.4

Expansion of the conditioning interval

The variables selected were now,

Set (ii) 14 64 68 97

These variables are illustarated in figure 6.5 where it will be seen that the only major difference is to replace the J wave by the H wave.

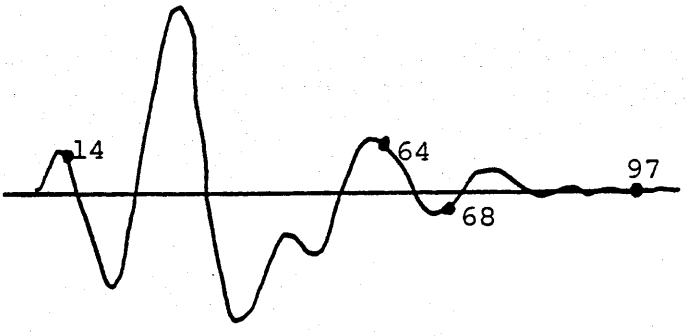


Figure 6.5

Selected features using an expanded conditioning interval

These variable sets are interesting in that the appear to have real meaning in terms of the data being analysed but one would wish to ensure that they actually performed well as part of a complete discriminatory procedure. In order to assess their relative performance they were compared with four other sets of variables, namely

Set (iii) 14 30 56 62 68 97

This is an amalgamation of (i) and (ii) using both the H and

J waves.

Set (iv) 8 16 24 32 96

This set consists of every eighth variable from the length of the wave.

Set (v) 5 12 18 28 40 48 54 61 69 78 93

Selected as having the largest absolute differences between the means of the two (transformed) classes.

Set (v) 10 30 42 46 59 68 76 82 97

Selected using a forward sequential method and the F-statistic commonly employed with normal data.

Now to judge the comparitive merits of these data sets one needs both a classifier and a measure of performance and, of course, the selection is vast. Our choice of classifier is defined as that which maximises the measure of separation, R, that was used earlier to select the variables in sets (i) and (ii). Thus we seek the linear combination of the variables that maximises the measure R. There is no guarantee that this method will be equally suitable for each data and one must be careful when interpreting the results. What is more we have used the apparent error rate to measure performance, which as we know is likely to be falsely optimistic. However we are interested in differences and not

absolute performance and thus the bias is less important. Table 6.8 shows the results of the comparison.

Table 6.8

A comparison of six sets of variables

SET	APPARENT ERROR RATE
(i)	3.8%
(ii)	2.3%
(iii)	.0%
(iv)	9.9%
(v)	9.9%
(vi)	10.7%

Even with all of our provisos this still seems suggest that our variables are preferable.

In order to get some feel for the absolute performance of these data sets, numbers (i), (ii) and (iii) with 4, 5 and 6 variables were analysed in an experiment whereby 30 pathological and 41 normal cases were used to define the classifier and the rest were used to test it. The results of the averages of four runs are shown in figure 6.6. This graph also shows the results obtained by Hanka(1978) when he analysed the same data set. On that occassion the variables were selected using the Mahalanobis distance and subsequent classification was by hyperquadrics. This is the type of

analysis that one would perform were the data multivariate normal. Once again the features selected by the new method seem to perform well.

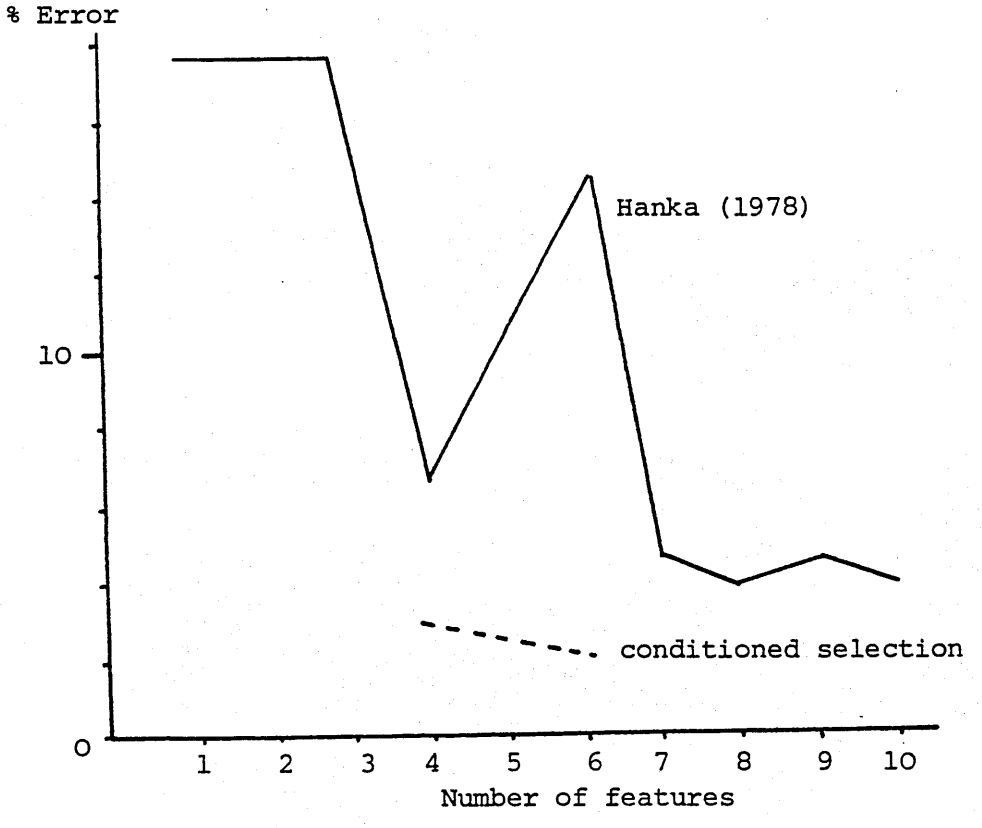


Figure 6.6
A comparison of conditioned selection with the results of Hanka (1978)

We have come up against all of the problems associated with testing feature selection methods and must be very careful when drawing conclusions that we are aware of the difficulty of disentangling the selection from all of the other factors in the analysis. All in all, however, the variables selected are meaningful and appear to perform well suggesting that the method may have wider application.

7 THE USE OF PRINCIPAL CO-ORDINATES FOR FEATURE EXTRACTION

7.1 Introduction

We saw in chapter five how the most commonly used approaches to transformed feature extraction are the linear methods derived from principal components or canonical variate analysis. In this chapter we consider another well established technique for multivariate analysis and investigate its potential as a method of feature extraction.

Principal co-ordinate analysis was devised under the name classical scaling by Torgensen(1952). It is essentially a metric version of the popular non-metric scaling techniques already in common use for feature extraction. Further it has the very desirable property that it includes both principal component analysis and canonical variate analysis as special cases. It will be seen that principal co-ordinate analysis is both versatile and produces useful features.

Principal co-ordinate analysis only requires that we should be able to specify a suitable distance matrix for the

data and is thus able to cope with any set of data for which a sensible metric or measure of similarity can be defined. This is important since it is often far easier to define a distance matrix than to form a probability model, such as when for example, there are mixtures of discrete and continuous variables.

The major problem when analysing a distance matrix is its size which in realistic problems can be extremely large. We therefore consider ways in which the computational load may be reduced whilst still analysing the whole data set.

7.2 Principal Co-ordinates

In this section we will outline the theory of principal co-ordinate analysis. Where details of proofs have been omitted they may be found either in the original articles, Torgensen(1952,1958), or in Gower(1966), or in any modern text book on multivariate analysis.

As its rather confusing name suggests principal co-ordinate analysis is very closely related to principal component analysis. It differs in that it seeks to analyse a distance matrix rather than a sums of squares matrix, but it does so in the same way, by looking at its eigenstructure. The object of the analysis is, by using the first few eigenvectors, to define a space which gives the best, in a least squares sense, representation of the original distances between points.

An investigation of a set of p- measurements on n subjects may proceed in one of two ways. Either one could make a, so called, Q-analysis of the nxn measures of association between the subjects, or one could make an R-analysis of the pxp measures of association between the variables. Clearly these two dual views give closely related results.

Suppose that the data collected has been mean centred and stored in a nxp matrix $\underline{X}$. One possible type of R-analysis would be to look at the eigenstructure of ,

$$\underline{W} = \underline{X}'\underline{X}$$

Here $\underline{W}$ will contain the within data sums of squares and products. An analysis of its eigenstructure would produce the usual principal components. We would obtain values λ_i and vectors $\underline{u}_i$ such that,

$$(\underline{X}'\underline{X})\ \underline{u}_i = \lambda_i\ \underline{u}_i$$

The corresponding Q-analysis would involve consideration of the eigenstructure of the matrix,

$$\underline{B} = \underline{X}\ \underline{X}'$$

Here $\underline{B}$ is an nxn matrix containing information about distances between subjects. Although $\underline{B}$ does not actually contain the distances themselves, as we shall see, it is a simple matter to obtain them from the contents of $\underline{B}$. An eigenstructure analysis of $\underline{B}$ would require us to find values μ_i and vectors $\underline{v}_i$ such that

$$\underline{B}\ \underline{v}_i = \mu_i\ \underline{v}_i$$

Such vectors $\underline{v}_i$ are known as the principal co-ordinates.

The connection between the two analyses becomes clear if we multiply the principal component equation by $\underline{X}$ to give,

$$\underline{X}\,(\underline{X}'\,\underline{X})\,\underline{u}_i = \lambda_i\,\underline{X}\,\underline{u}_i$$

or

$$(\underline{X}\,\underline{X}')\,\underline{X}\,\underline{u}_i = \lambda_i\,\underline{X}\,\underline{u}_i$$

Comparing this with the principal co-ordinate equation we see that,

$$\mu_i = \lambda_i$$

and that,

$$\underline{v}_i = K_i\,\underline{X}\,\underline{u}_i$$

If we normalise the eigenvectors so that,

$$\underline{u}_i'\underline{u}_i = 1$$

and

$$\underline{v}_i'\,\underline{v}_i = \lambda_i$$

then we will have,

$$\underline{v}_i'\,\underline{v}_i = K_i^2\,\underline{u}_i'\,\underline{X}'\,\underline{X}\,\underline{u}_i$$

or

$$\lambda_i = K_i^2 \lambda_i \underline{u}_i' \underline{u}_i$$

giving,

$$K_i = 1 \quad \text{and} \quad \underline{v}_i = \underline{X}\,\underline{u}_i$$

The consequence of the duality between principal co-ordinates and principal components is that we are dealing with the same set of eigenvalues. What is more, given any one set of eigenvectors it is a simple matter to obtain the other.

As we noted before $\underline{B}$ does not contain the distances themselves. However, the information that $\underline{B}$ does contain is equivalent to that contained in the matrix of Euclidean distances, $\underline{D}$, in the sense that starting with either it is possible to derive the elements of the other. The relationships that enable these calculations being,

$$d_{ij}^2 = b_{ii} + b_{jj} - 2b_{ij}$$

and

$$b_{ij} = -\tfrac{1}{2}(d_{ij}^2 - \bar{d}_{i.}^2 - \bar{d}_{.j} + \bar{d}_{..}^2)$$

where bars denote averages over the squared elements of $\underline{D}$.

Thus we may start with the Euclidean distances, derive $\underline{B}$ and then perform the principal co-ordinate analysis. Plotting the points with co-ordinates given by the rows of the vector $\mu_i \underline{V}_i$ we would reconstruct the original configuration. Since the distance information has been used we cannot be sure of the orientation of the points and so the solution may differ from the original configuration by a rotation, reflection or translation.

The ability to derive principal co-ordinates from a distance matrix opens up the possibility of using some other metric in place of the Euclidean distance and thus of progressing to a different solution. By this method we can analyse any set of data for which it is possible to define a sensible distance or similarity matrix.

If a non-Euclidean distance matrix is used as the starting point then problems may arise as the principal co-ordinate solution will try to reproduce the configuration of points in a Euclidean space and this may not be possible. However, it is usually the case that whilst an exact representation may not be possible a good approximation can be found. The non-Euclidean nature of the original configuration will show up in the eigenstructure analysis as a series of negative eigenvalues. Usually the main eigenvalues will be positive and only a few insignificant eigenvalues will be negative. Consequently a good approximation to the original configuration will be possible

in q-dimensional Euclidean space so long as the first q eigenvalues are positive and dominate the eigenstructure analysis.

We have already seen the duality between principal components and principal co-ordinates when a Euclidean distance is employed. Gower(1966) showed that another interesting duality exists between canonical variate analysis and principal co-ordinate analysis when one uses the Mahalanobis distance between k classes.

Suppose that we form the matrix $\underline{D}$ of Mahalanobis distances between classes, that is,

$$d_{ij} = (\bar{\underline{x}}_i - \bar{\underline{x}}_j)' \underline{W}^{-1} (\bar{\underline{x}}_i - \bar{\underline{x}}_j)$$

Using the relationships given above this implies that if we, without loss of generality, centre the means on zero, we find that,

$$b_{ij} = \bar{\underline{x}}_i' \underline{W}^{-1} \bar{\underline{x}}_j$$

and if the means are stored in a kxp matrix $\underline{X}$ then we will have,

$$\underline{B} = \bar{\underline{X}} \underline{W}^{-1} \bar{\underline{X}}'$$

Now it will always be possible to factorise $\underline{W}^{-1}$ in the form,

$$\underline{W}^{-1} = \underline{U}\,\underline{U}'$$

so that we may write,

$$\underline{B} = (\overline{\underline{X}}\,\underline{U})\,(\overline{\underline{X}}\,\underline{U})'$$

Hence an eigenstructure analysis of $\underline{B}$ will be the dual of an eigenstructure analysis of,

$$(\overline{\underline{X}}\,\underline{U})'(\overline{\underline{X}}\,\underline{U})$$

which would involve the solution of,

$$(\overline{\underline{X}}\,\underline{U})'(\overline{\underline{X}}\,\underline{U})\,\underline{u}_i - \lambda_i\,\underline{u}_i = 0$$

which rearranges to,

$$\underline{U}'\left[\overline{\underline{X}}'\,\overline{\underline{X}} - \lambda_i\,\underline{U}'^{-1}\,\underline{U}^{-1}\right]\underline{U}\,\underline{u}_i = 0$$

or

$$\underline{U}'\left[\overline{\underline{X}}'\,\overline{\underline{X}} - \lambda_i\,\underline{W}\right]\underline{U}\,\underline{u}_i = 0$$

and this is exactly the form required for a canonical variate analysis.

7.3 Adding a point to a Principal co-ordinate analysis

Having obtained a principal co-ordinate representation of the training sets using some appropriately dimensioned space, one might wish to be able to add an unclassified point to see which class it is most likely to have come from. The procedure for doing this was given by Gower(1968), although not in the context of discrimination.

Suppose that one has analysed $\underline{B}=\underline{X}\underline{X}'$ based on n subjects and that one now wishes to find the co-ordinates of an n+1st point in that same space. Let the eigenvectors $\underline{v}_k$ of $\underline{B}$ be normalised so that

$$\underline{v}_k' \underline{v}_k = \lambda_k$$

and suppose that we have also got the distances from the new point to the n original points,

$$d^2_{i,n+1} = \sum_{k=1}^{p} (V_{n+1,k} - V_{i,k})^2 \qquad i = 1, \ldots, n$$

From this information we may locate the new point relative to the principal co-ordinate axes. It may happen that this will require an extra dimension but that is more of a theoretical problem than a practical one, and in any case there is sufficient information to solve for the extra

dimension.

A little algebra on the distance expression gives,

$$d^2_{i,n+1} = d^2_{n+1} + d^2_i - 2 \sum_{k=1}^{p} v_{i,k}\, v_{n+1,k} \qquad i = 1, \ldots, n$$

where d_i is the distance from the origin. Because the data is mean centred,

$$\sum_{i=1}^{n} d^2_{i,n+1} = n\, d^2_{n+1} + \sum_{i=1}^{n} d^2_i$$

so that

$$2 \sum_{k=1}^{p} v_{i,k}\, v_{n+1,k} = d^2_{i,n+1} - \frac{1}{n} \sum_{i=1}^{n} (d^2_i - d^2_{i,n+1})$$

This is best represented in matrix notation, so that if we take,

$\underline{U}$ = (nxn) matrix of ones

$\underline{d}$ = (nx1) vector of elements $d^2_i - d^2_{i,n+1}$

$\underline{v}_{n+1}$ = (px1) vector of required co-ordinates

$\underline{v}$ = (nxp) matrix of the co-ordinates of the original data

The equation then takes the form,

$$2\underline{V}\,\underline{v}_{n+1} = \underline{d} - \tfrac{1}{2}\,\underline{U}\,\underline{d}$$

which gives the required solution,

$$\underline{v}_{n+1} = \tfrac{1}{2}(\underline{V}'\,\underline{V})^{-1}\,\underline{V}'\,\underline{d}$$

7.4 An analysis of the Psychiatric data

This data set is described fully in chapter three, but briefly, consists of 146 subjects classified as having one of five types of depression. The subjects were all rated according to 42 symptoms and we assume to start with that it is reasonable to calculate Euclidean distances between subjects on the basis of these symptom ratings.

Rather than analyse the 146 subjects we have taken the distances between class means. Since Euclidean distances have been used we have the equivalent of a principal component analysis of the class means. There are five classes and hence four non-zero eigenvalues as shown in table 7.1. From these eigenvalues it will be seen that the class means can be well represented in a plane, since this explains 86% of the variablity. The class means are plotted in figure 7.1 and show clearly the separation available from these two features.

Table 7.1

Euclidean distance analysis of class means

Eigenvalues	50.79	20.21	6.05	5.42
	61.6%	24.5%	7.3%	6.6%
Eigenvectors	-3.60	-2.61	1.06	-0.32
	-1.87	.46	-0.96	1.76
	-1.31	1.16	-1.32	-1.49
	1.20	2.87	1.48	0.04
	5.58	-1.88	-0.26	0.02

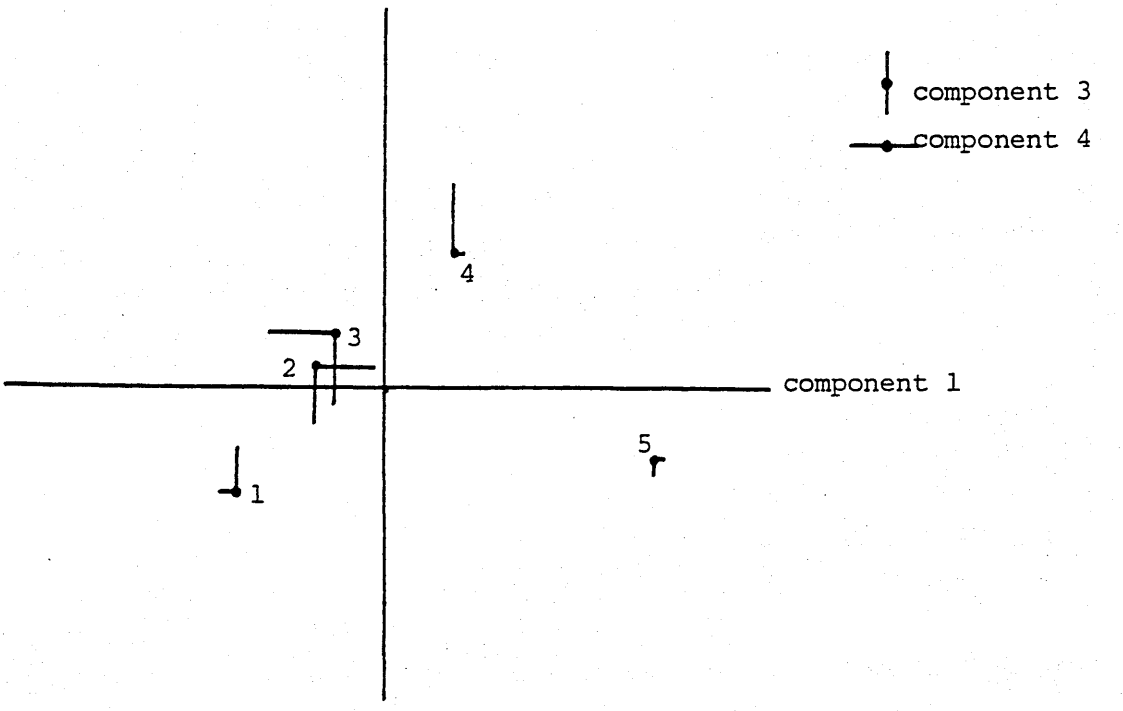


Figure 7.1

Group means from the psychiatric data

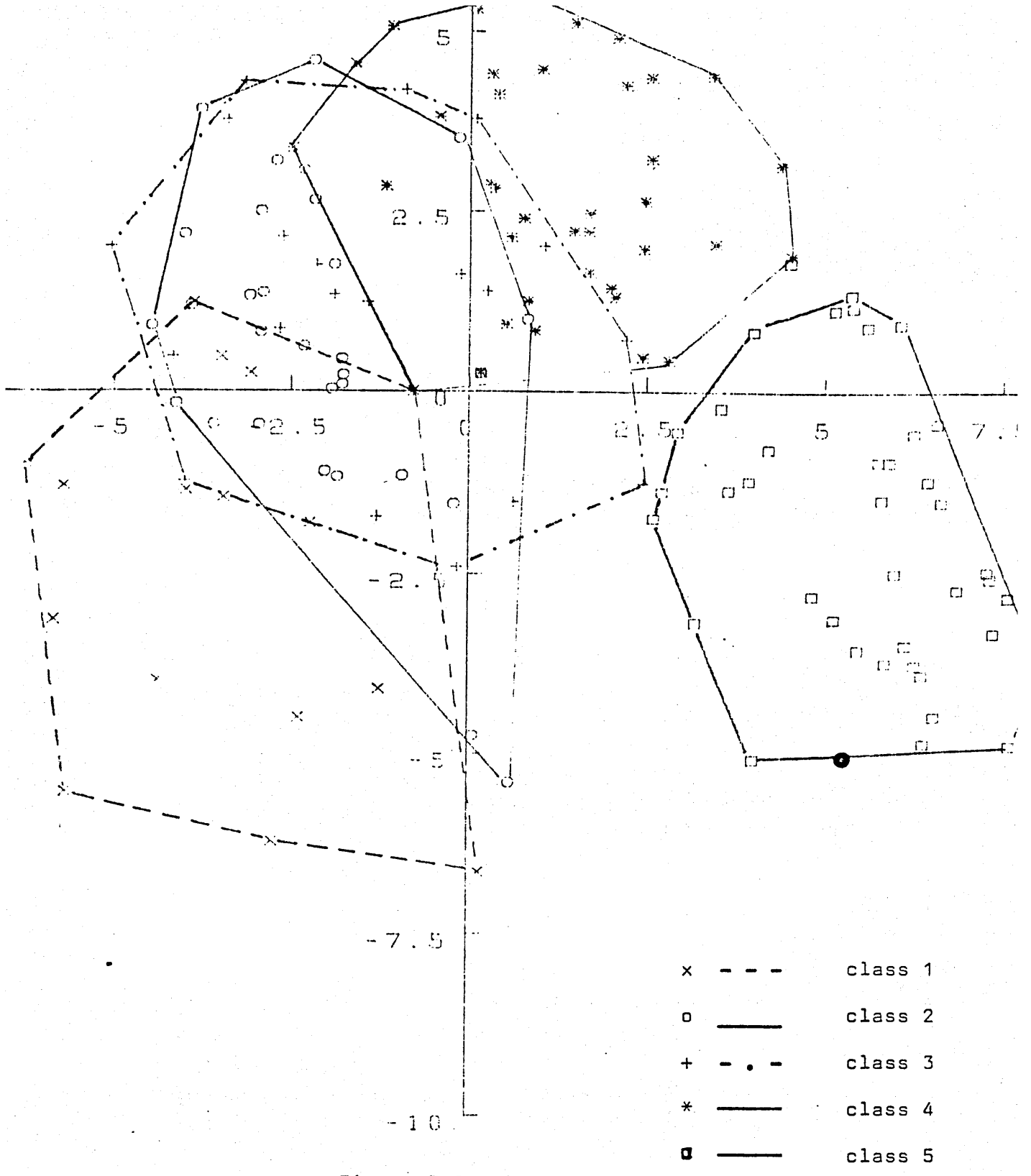


Figure 7.2

PCA based on Euclidean distances and showing individual cases

Having obtained this solution we can then add the individuals to the plot using the methods desribed in section 7.3. This gives the picture presented as figure 7.2. We can now see that if we omit the one case which appears to have been misclassified, (shown in blue in figure 7.2), then classes 1,4 and 5 are quite well separated in just two dimensions, but classes 2 and 3 overlap with each other and with classes 1 and 4. Not only are classes 2 and 3 close together but they are also the least well represented in two dimensions, as we can tell by reference to table 7.1 were classes 2 and 3 are seen to have the largest components in the third and fourth dimensions. This is not a co-incidence, for eigenstructure analysis is a form of least squares and so tends towards a better representation of large values. It should be noted that quite different results would have been obtained had we performed the analysis on the 146 subjects. The analysis presented here is specifically geared towards emphasising differences between classes.

As mentioned before the great advantage of the principal co-ordinate analysis is that it does not require us to stick to Euclidean distances. This is of particular advantage with data sets such as these where one has doubts about the appropriateness of the distance measure.

An alternative measure of distance that does not rely on the numerical scoring of the categories would provide an

indirect validation of the scoring. In order to define a distance measure without using the scoring we have taken, for each class, the pattern of most commonly given symptom categories. Thus one class might have category 2 as the most common reply to symptom 1, category 4 as the most common reply to symptom 2, and so on. Treating this profile as the definition of the centre, the distances between classes was then taken as the proportion of times that the two class centres are different. Similarly the distance of a subject from a class centre is the proportion of times that the subject does not have the same response as the centre.

The eigenstructure of the distances between class centres is shown in table 7.2 and the first two components will be seen to account for 84% of the variability. These two components are plotted in figure 7.2.

Table 7.2

Non-Euclidean distance principal co-ordinate analysis

Eigenvalues	.19	.08	.04	.00	-.01
	59.4%	25.0%	12.5%		
Eigenvectors	-0.25	0.04	0.12		
	-0.13	0.08	-0.05		
	-0.04	-0.03	-0.14		
	0.11	-0.24	0.04		
	0.31	0.14	0.04		

The distances cannot be represented completely even in a five dimensional Euclidean space, however the two dimensional approximation is a good one. Allowing for the fact that these configurations may be rotated, this representation is very similar to that obtained previously and shown in figure 7.1.

The subjects have been added to the two dimensional representation and the results are shown as figure 7.4. Once again the misclassified cases shows up clearly althoughthe overall distinction between classes is less well defined. It would appear that the scoring system suggested by the psychiatrist is reasonably good as a basis for class definition.

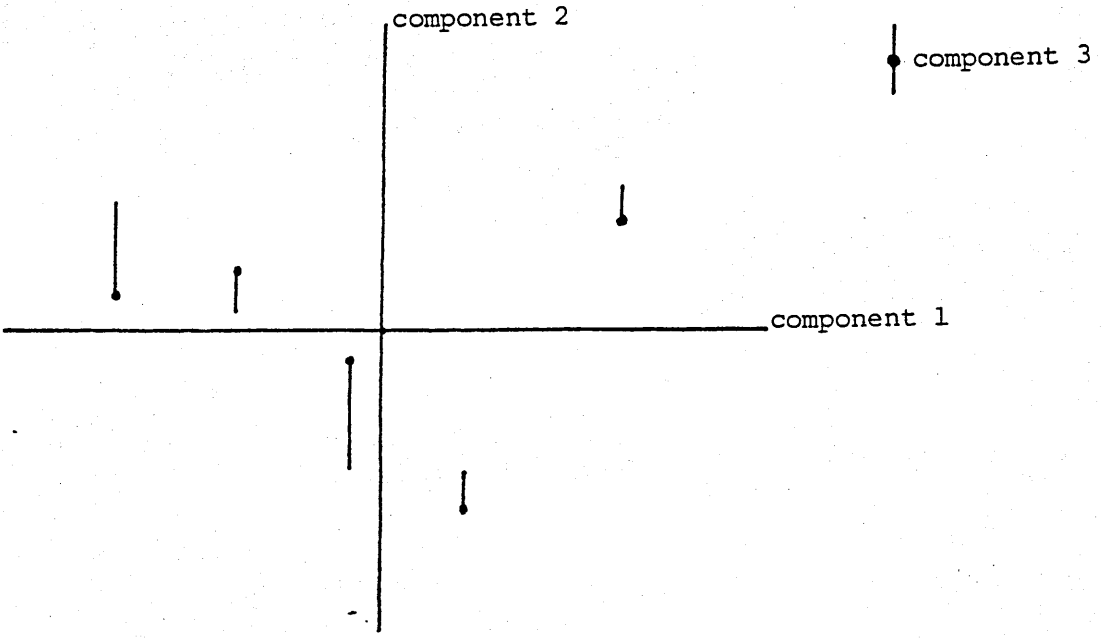


Figure 7.3

Group centres from the psychiatric data

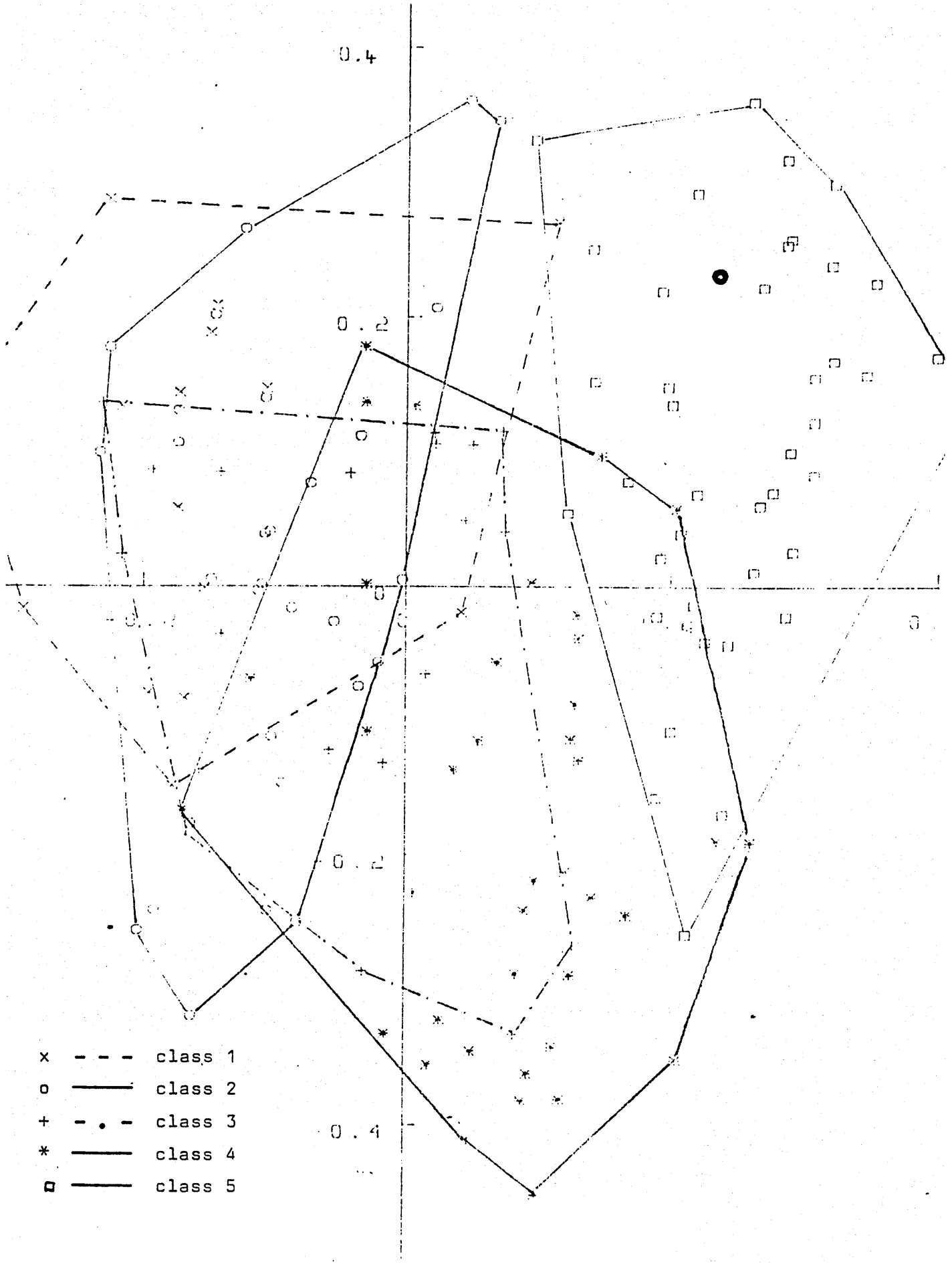


Figure 7.4

PCA based on a non-Euclidean distance and showing individual cases

7.5 Weighted analysis

One way of thinking of an eigenstructure analysis is as a matrix factorisation. Thus if $\underline{B}$ has eigenvalues n_i and normalised eigenvectors $\underline{v}_i$ then,

$$\underline{B} = \underline{v}_1 \underline{v}_1' + \underline{v}_2 \underline{v}_2' + \ldots + \underline{v}_p \underline{v}_p'$$

Assuming that these eigenvectors are numbered in the order of the size of their eigenvalues then we may make a rank r approximation to $\underline{B}$ by taking the first r terms,

$$\underline{B}_r = \underline{v}_1 \underline{v}_1' + \ldots + \underline{v}_r \underline{v}_r'$$

Eckart and Young(1936) first showed that this is the best rank r approximation to $\underline{B}$ in the least squares sense. That is it minimises,

$$|| \underline{B} - \underline{B}_r ||_F$$

where

$$|| \underline{A} ||_F = \sqrt{\sum_i \sum_j a_{ij}^2}$$

is the usual Frobenius norm.

It would be useful if we could weight this analysis so

that certain of the distances are better reproduced than others. That is, it would be nice to be able to minimise,

$$\sum_i \sum_j w_{ij} (b_{ij} - b_{rij})^2$$

This is a solvable problem and Gabriel and Zamir(1979) have compared different iterative algorithms for the minimisation. The computation involved in the full analysis is very long but there is one special case that can be solved much more simply. Yet this special case still enables us to go some way towards weighting the analysis.

Suppose that the weights can be factorised so that,

$$w_{ij} = s_i s_j$$

In this case the weight given depends upon the product of two terms each depending on one of the classes or subjects in the analysis. Minimising the weighted least squares expression,

$$\sum_i \sum_j s_i s_j (b_{ij} - b_{rij})^2$$

is equivalent to minimising,

$$|| \underline{S} (\underline{B} - \underline{B}_r) \underline{S} ||_F$$

where $\underline{S}$ is a diagnonal matrix with elements s_i.

Now to minimise the quantity all we need to do is to

minimise,

$$||\underline{S}\,\underline{B}\,\underline{S} - \underline{A}_r||_F$$

from which we can derive $\underline{B}_r$ from,

$$\underline{B}_r = \underline{S}^{-1}\,\underline{A}_r\,\underline{S}^{-1}$$

The procedure for finding $\underline{B}_r$ is thus to first form the usual least squares fit to $\underline{S}\underline{B}\underline{S}$, which may be obtained by an eigenstructure analysis of that matrix.

Although we have placed a severe constraint upon the weights w_{ij}, this restricted analysis can still be very useful. We might, for example, when analysing the distances between classes, take

$$S_i = \sqrt{n_i}$$

in which case the diagonals of $\underline{B}$ will be weighted according to sample size. The tendency will then be to improve the fit to large classes at the expense of the smaller ones. Obviously other weighting schemes are possible if one has specific interest in differences between particular classes.

Returning to the psychiatric data the class sizes are respectively 15,31,20,39 and 41. Using the Euclidean distance and weights calculated from the square roots of the sample sizes we obtain the solution shown in table 7.3 and figure 7.6. clearly the solution differs very little from the original analysis.

Table 7.3

Euclidean distance analysis weighted by class size

Lengths	46.69	20.31	6.38	6.13
	58.7%	25.5%	8.0%	7.7%
Vectors	-3.03	2.87	1.63	1.00
	-1.98	0.15	-1.86	0.56
	-1.49	-0.58	0.02	-2.12
	0.67	-3.32	0.42	0.54
	5.83	0.87	-0.21	0.01

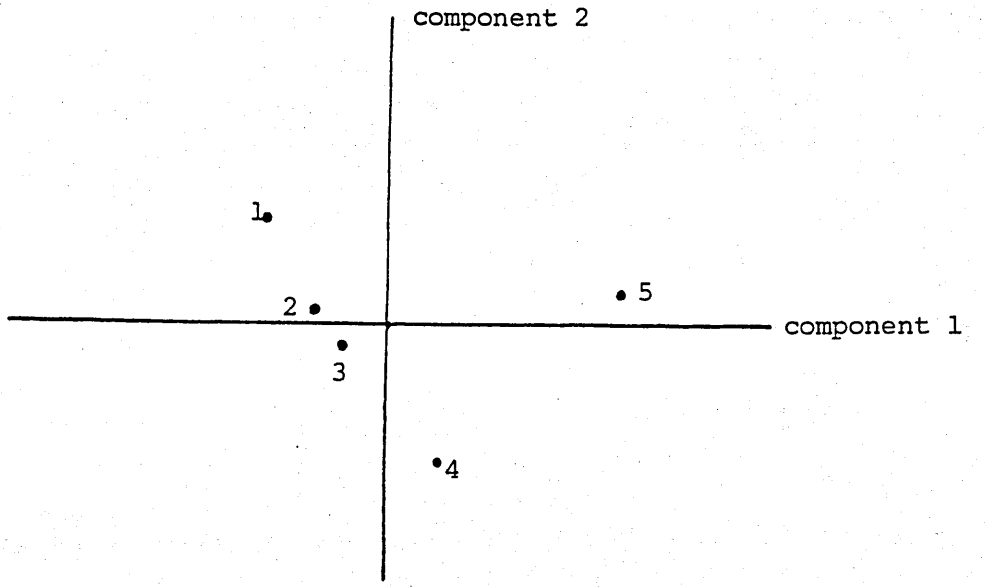


Figure 7.5

Weight analysis of the class means

We might however wish to find a space in which the distinction between classes 2 and 3 is emphasised possibly at the expense of the other classes. Thus we would wish to choose s_i so that s_2 and s_3 are largest. In the analysis that follows the values chosen for s_i were $(1,\sqrt{10},\sqrt{10},1,1)$.

Table 7.4 shows the results of the analysis with the eigenvectors given in the order of the eigenvectors of $\underline{S}\underline{B}\underline{S}$. The reason for this being that it is those components that emphasise the differences that we are interested in even though they do not account for the largest proportions of the overall variation.

Table 7.4

Euclidean distance analysis weighted to emphasise classes 2 and 3

Lengths	41.76	7.94	21.58	11.18
	50.6%	9.6%	26.2%	13.6%
	-1.99	1.25	3.93	-0.28
	-2.31	1.40	-0.60	0.24
	-1.81	-1.92	-0.09	0.28
	0.77	-0.85	-2.14	-2.46
	5.35	0.12	-1.14	2.22

The class centres are shown in figure 7.6 and the individual subjects have been added in figure 7.7. Clearly

the approach separates clsses 2 and 3 to a greater extent than previously.

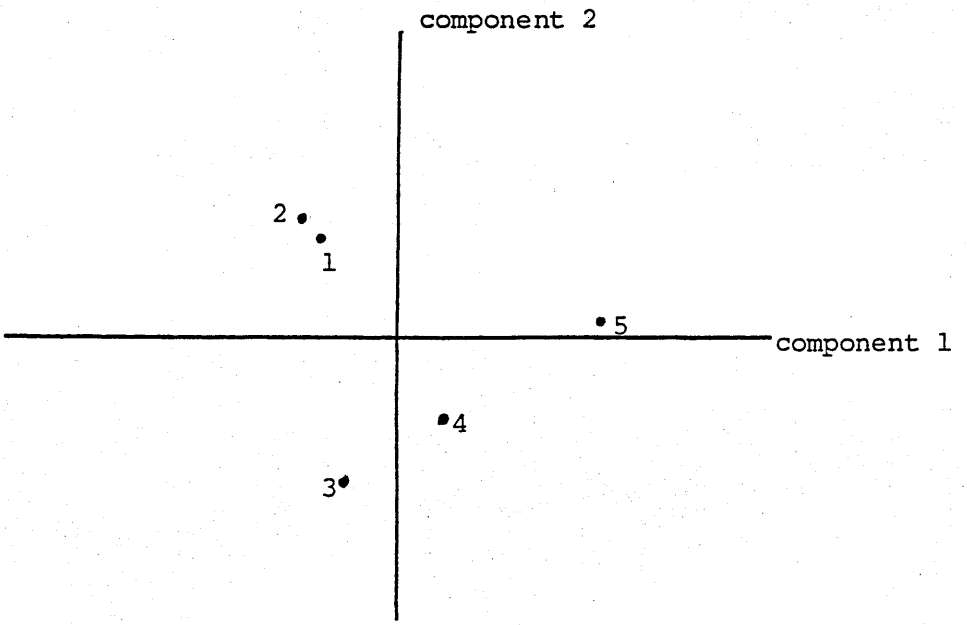


Figure 7.6

Weights chosen to separate classes 2 and 3

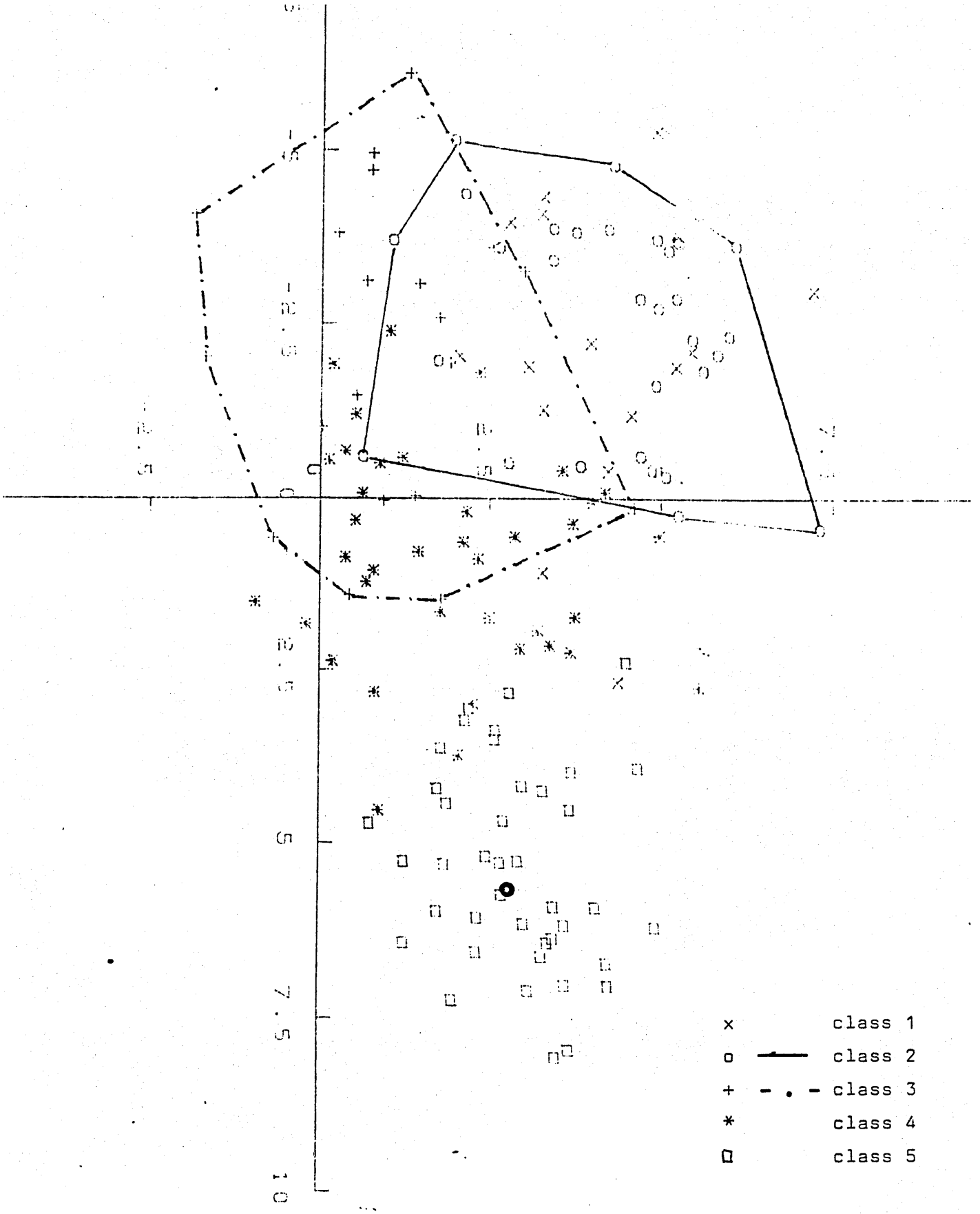


Figure 7.7

A weighted PCA intended to distinguish classes 2 & 3

7.6 An analysis of the Bcg's

When the number of classes is small the idea of looking at the separation of their centres will no longer be useful. In order to show how this problem may be overcome we will now analyse the data on the ballistocardiograms. Thus we have two training sets consisting of 81 normal and 50 pathological waves. In this analysis we have used the amplitudes and locations of the H,I,J,K,L,M and N waves as described in section 4.2.

Since there are only two classes others need to be generated before a sensible analysis can be performed. To this end a cluster analysis was performed on each class. The furthest neighbour hierarchical method was used since this tends to produce compact clusters, rather than the long strings of points that can result from nearest neighbour analyses. As a result four fairly well defined clusters were identified amongst the normal cases and three amongst the pathological.

There is little doubt that the search for clusters was simpler in this example by the nature of the data, consisting as it does of a series of repeat measurements on a smaller number of individuals. However this stage should

not be critical to the performance of the whole analysis and any reasonable subdivision would suffice.

The result of the clustering is then to produce seven classes, the four normal being denoted by N_1, N_2, N_3, and N_4 whilst the three pathological will be denoted by P_1, P_2 and P_3. A Euclidean distance matrix was constructed for these seven classes using their class means, and this was then subjected to a principal co-ordinate analysis with equal weights to produce the results shown in table 7.5 and illustrated in figure 7.8. It will be noted that the second pathological cluster is a little nearer to the normal classes than the other two although this sub-class would have moved further away in the third dimension.

Table 7.5

P.C.A. of the Bcg's

Lengths	16817	1768	965	356	105	9	0
Vectors	-34	1	-6				
	-37	-10	-2				
	-13	-17	-5				
	-49	23	-10				
	37	-22	4				
	-6	12	27				
	102	14	-8				

Figure 7.9 shows the results of adding individuals and confirms that the separation between normal and pathological is indeed good, especially bearing in mind the fact that we are only using two dimensions.

Our analysis has resulted in one pathological subclass that is close to the normals and two that are well separated. The obvious question is then, could we increase the separation between the normals and the close pathological sub-class perhaps at the expense of some of the separation between the normals and the remoter pathological subclasses?

In order to achieve this one needs to give greater weight to N_1, N_2, N_3 and P_2. The weights chosen were $(\sqrt{2},\sqrt{2},\sqrt{2},1,1,\sqrt{2},1)$. These small changes in the weights produced a marked effect on the results as shown in table 7.6 and figure 7.10. When the individuals are added as they are in figure 7.11 we can see that all but a handful of cases are now well separated.

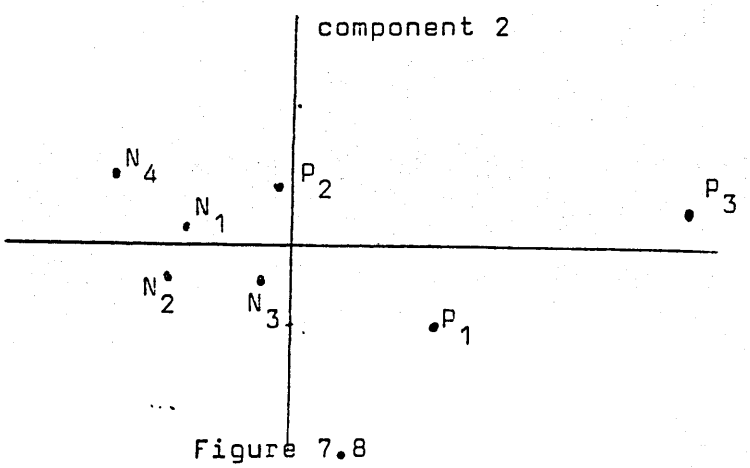


Figure 7.8

Class centres of the seven Bcg clusters

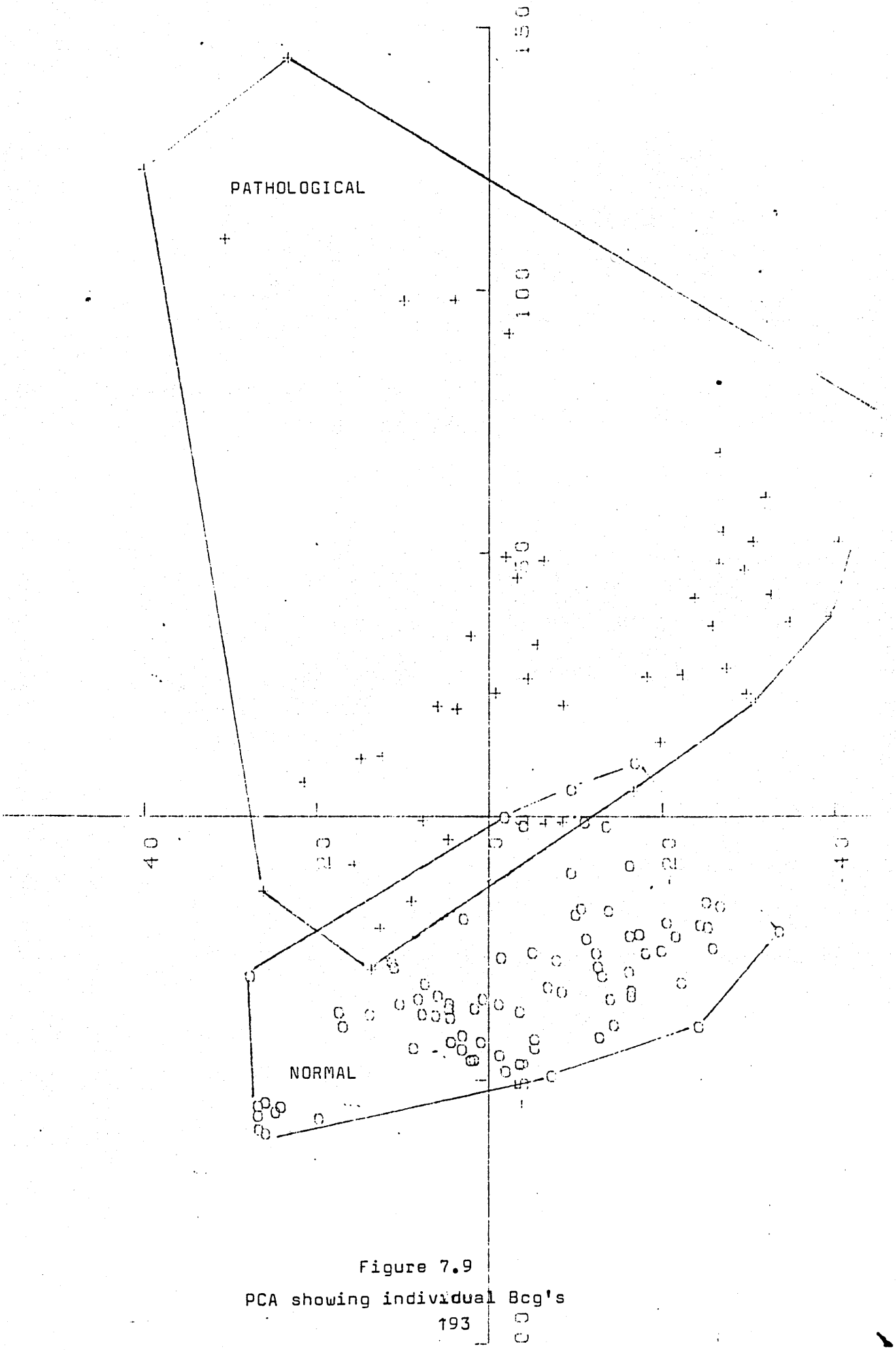


Figure 7.9

PCA showing individual Bcg's

Table 7.6

Weighted P.C.A. of the Bcg's

Lengths	16806	1225	1470	367	112	10	0
Vectors	-34	-4	5				
	-37	-8	-5				
	-13	-14	-12				
	-49	6	25				
	36	-11	-20				
	-6	29	-7				
	102	2	14				

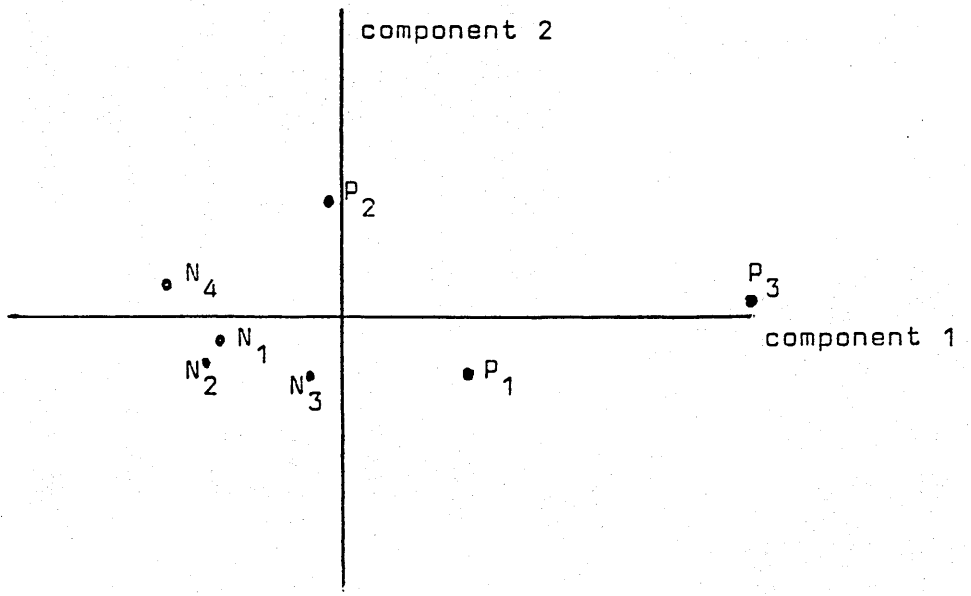


Figure7.10

Weighted PCA of the Bcg's

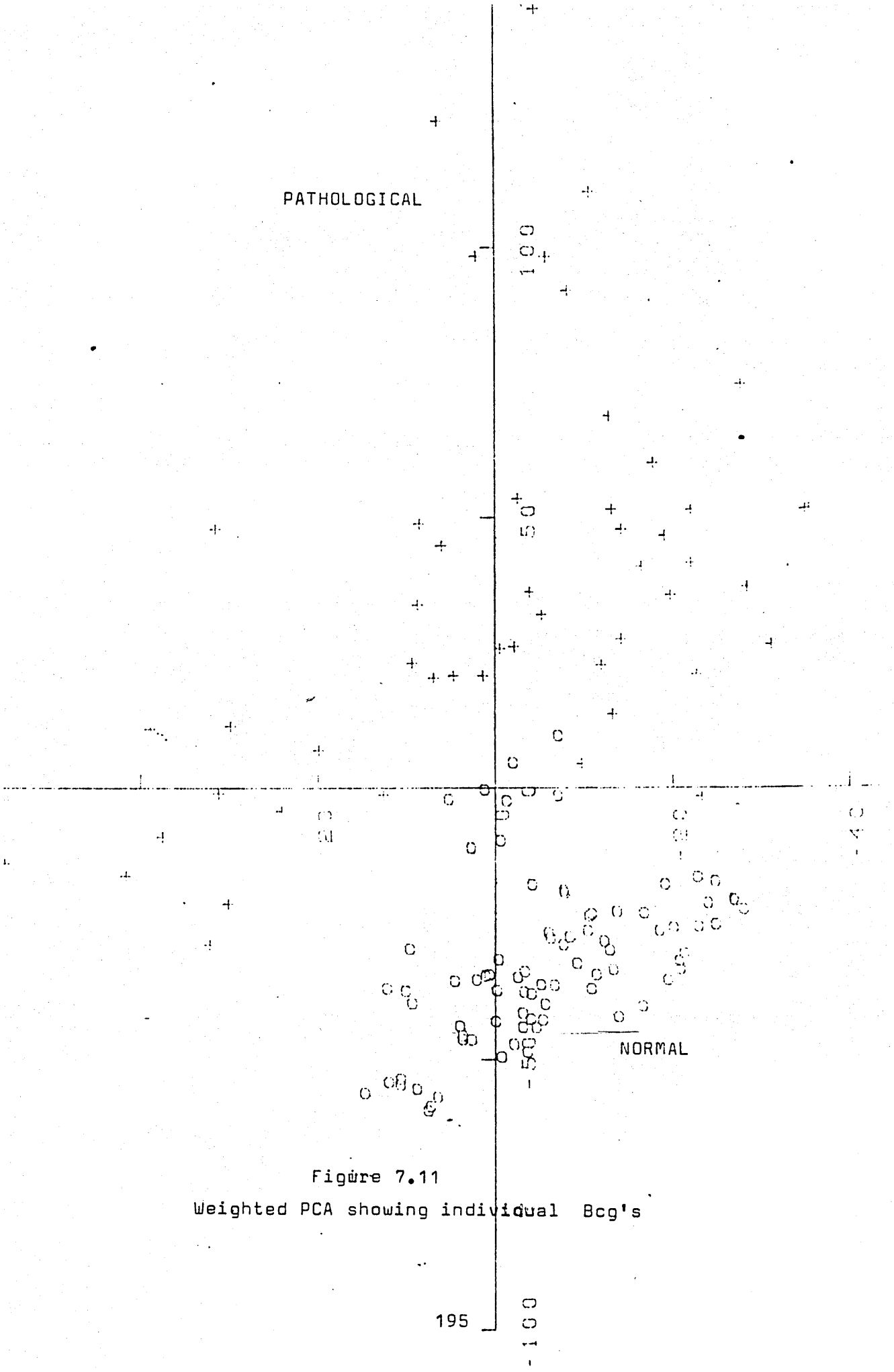


Figure 7.11
Weighted PCA showing individual Bcg's

8 LINEAR CLASSIFIER DESIGN

8.1 Introduction

Having considered some new methods of feature selection we will now investigate the second stage in the analysis, namely the design of the classifier. Clearly the design of the classifier ought to be linked to the method of feature selection and in subsequent chapters we will study the properties of a linear classifier specifically designed to be used will the non-parametric approach to feature selection that was described in chapter six.

First however, we review the currently available methods for defining a classification rule based on linear functions of the selected features. We will restrict our review by only considering the most common situation, namely that in which one has cases previously classified as coming from one of the classes. Obviously there is some overlap between linear feature selection, considered previously, and linear classifier design, for given a linear feature one merely has to add class boundaries in order to produce a linear classifier.

Linear classification has receieved a considerable amount of attention since Fisher's 1936 paper. Not only is it optimal for normally distributed classes with equal covariance structures but it has been found to be very robust, simple to estimate and easy to interpret. Even if as Duda and Hart(1973) observe,

' far more papers have been written about them than the subject deserves'

it is as well to remember that linear classifiers have performed well in many applications.

A linear classifier is then, for the two class problem, a combination of a linear function $g(\underline{x})$ of the p observed values $\underline{x}$, and a boundary value k such that,

if $g(\underline{x}) > k$ $\underline{x}$ is classified as w_1

if $g(\underline{x}) < k$ $\underline{x}$ is classified as w_2

The linearity restriction is not as constraining as it might at first appear for we could generate r new variables,

$$\phi_i(\underline{x}) \qquad\qquad i = 1, \ldots, r$$

from the p original variables and then form a linear combination $g(\phi_i(\underline{x}))$. This idea is most commonly employed when ϕ is taken to contain the squares and crossproducts of the original variables, the result being that ϕ is a hyperquadric in the original data $\underline{x}$. Such classifiers based on functions ϕ are sometimes known as generalised linear

discriminant functions.

Another way in which linear functions may be used is in combination. So that a curved boundary could be approximated by a piecewise linear function as shown in figure 8.1

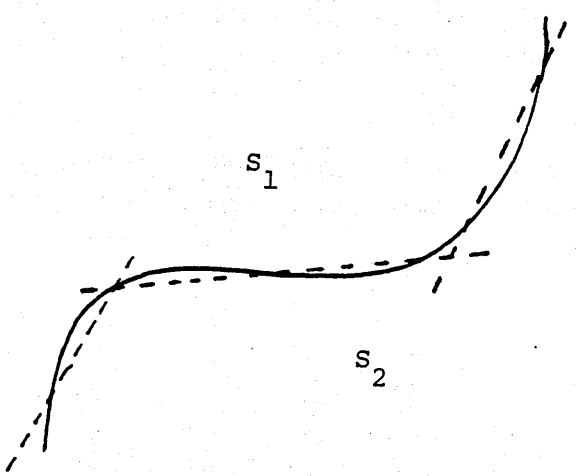


Figure 8.1
A piece-wise linear function

Although we concentrate on the two class problem all of our results may be generalised to cope with multiple classes, either by defining extra boundary values k_i as in figure 8.2, or by defining more than one linear function as in figure 8.3.

8.2 Bayes Classifcation of Normal Data

As we mentioned in the introduction to this chapter, much of the impetus for the use of linear classifiers has come about as a result of their optimality for pairs of normally distributed classes with equal covariance structures. To show the Bayes classifier under such circumstances is linear one only needs to consider the equations of their densities. Thus suppose that we have two p-dimensional normal distributions with means $\underline{\mu}_1$ and $\underline{\mu}_2$ and common covariance matrix $\underline{\Sigma}$. Further let the classes w_1 and w_2 have prior probabilities $P(w_1)$ and $P(w_2)$.

Bayes classification is achieved by assigning $\underline{x}$ to w_1 if,

$$f(\underline{x}|w_1)\ P(w_1) > f(x|w_2)\ P(w_2)$$

In the case of multivariate normal data the densities have the form,

$$f(\underline{x}|w_j) = |2\pi\ \underline{\Sigma}|^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}(\underline{x} - \underline{\mu}_j)'\ \underline{\Sigma}^{-1}\ (\underline{x} - \underline{\mu}_j)\}$$

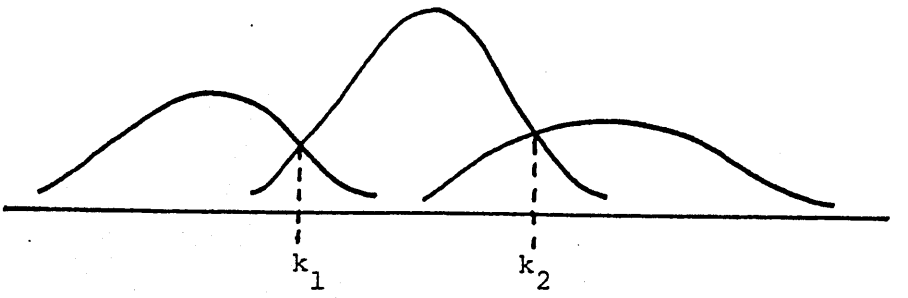


Figure 8.2

Linear classification with two constants

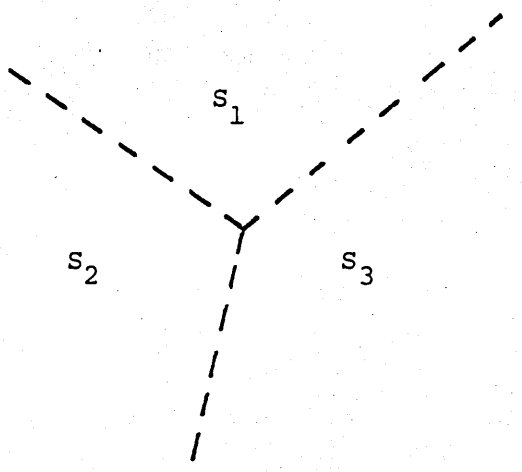


Figure 8.3

Linear classification by multiple functions

which have the form,

$$(\underline{x} - \underline{\mu}_j)' \, \underline{\Sigma}^{-1} \, (\underline{x} - \underline{\mu}_j) \quad = \quad \text{constant}$$

This family of curves is a set of ellipses centred on $\underline{\mu}_j$ and having the same orientation for both distributions, as in figure 8.4.

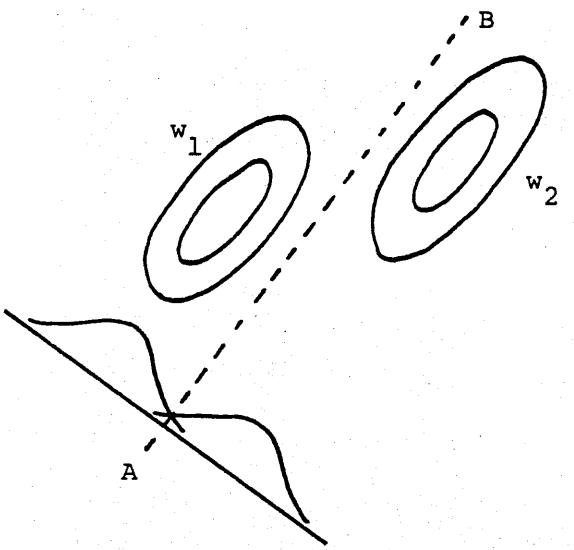


Figure 8.4
Linear discrimination between two bivariate normals

The boundary for the Bayes classifier is the locus of the points $\underline{x}$ for which,

$$(\underline{x} - \underline{\mu}_1)' \, \underline{\Sigma}^{-1} \, (\underline{x} - \underline{\mu}_1) \quad = \quad (\underline{x} - \underline{\mu}_2)' \, \underline{\mu}^{-1} \, (\underline{x} - \underline{\mu}_2)$$

and is shown by the line AB. We may also think of $g(\underline{x})$ as a function projecting the observation $\underline{x}$ onto a line normal to that boundary. Because linear functions of normal variables are themselves normal the resulting distributions will also

Consequently, on taking logs one obtains,

$$\underline{x}' \underline{\Sigma}^{-1}(\underline{\mu}_1 - \underline{\mu}_2) - \tfrac{1}{2}(\underline{\mu}_1 + \underline{\mu}_2)' \underline{\Sigma}^{-1} (\underline{\mu}_1 - \underline{\mu}_2) > \ln \{ \frac{P(w_2)}{P(w_1)} \}$$

The Bayes classifier is thus linear in $\underline{x}$, with

$$g(\underline{x}) = \underline{x}' \underline{\Sigma}^{-1} (\underline{\mu}_1 - \underline{\mu}_2)$$

and

$$k = \tfrac{1}{2}(\underline{\mu}_1 + \underline{\mu}_2)' \underline{\Sigma}^{-1} (\underline{\mu}_1 - \underline{\mu}_2) + \ln \{ \frac{P(w_2)}{P(w_1)} \}$$

Unfortunately when the covariance structures differ, say taking values $\underline{\Sigma}_1$ and $\underline{\Sigma}_2$, then the Bayes classifier is no longer linear. In fact $\underline{x}$ would be classified as coming from w_1 if,

$$\tfrac{1}{2}\ln \left| \frac{\underline{\Sigma}_1}{\underline{\Sigma}_2} \right| - \tfrac{1}{2}(\underline{x} - \underline{\mu}_1)' \underline{\Sigma}_1^{-1} (\underline{x} - \underline{\mu}_1) + \tfrac{1}{2}(\underline{x} - \underline{\mu}_2)' \underline{\Sigma}^{-1} (\underline{x} - \underline{\mu}_2) > \ln \{ \frac{P(}{P(}$$

So that the class boundary is a p-dimensional hyperquadric.

It is both simple and instructive to visualise the linear classifier when p=2 and the prior probabilities are equal. In this case the distributions are described by their contours of equal density,

$$f(\underline{x}|w_j) = \text{constant}$$

as pictured in the figure. All of these ideas extend directly into p-dimensions.

In this simple case we may derive the means and variances of the transformed values without difficulty so that,

$$E\{g(\underline{x}) \mid w_j\} = \underline{\mu}_j' \, \underline{\Sigma}^{-1} \, (\underline{\mu}_1 - \underline{\mu}_2)$$

and

$$\begin{aligned} \text{Var}\{g(\underline{x}) \mid w_j\} &= (\underline{\mu}_1 - \underline{\mu}_2)' \, \underline{\Sigma}^{-1} \, (\underline{\mu}_1 - \underline{\mu}_2) \\ &= D^2 \end{aligned}$$

Thus the two univariate distributions have common variance D^2 and their means differ by D^2.

In the case where the prior probabilities are equal, the constant k will lie mid-way between the two class means and the resulting probabiliy of error will be,

$$\Phi\ (-\tfrac{1}{2}D)$$

Unfortunately it is rare to be in the position of knowing $\underline{\mu}_j$ and $\underline{\Sigma}$ and one therefore needs to estimate $g(\underline{x})$ and k. This is usually done by taking estimates $\hat{\underline{\mu}}_j$ and $\hat{\underline{\Sigma}}$ of the parameters and plugging them into the formula for the Bayes classifier. The most popular such estimates being those based on the maximum likelihood. Thus we obtain

$$\hat{g}(\underline{x}) = \underline{x}' \, \underline{S}^{-1} \, (\bar{\underline{x}}_1 - \bar{\underline{x}}_2)$$

and

$$\hat{k} = \tfrac{1}{2}(\bar{\underline{x}}_1 + \bar{\underline{x}}_2)' \, \underline{S}^{-1} \, (\bar{\underline{x}}_1 - \bar{\underline{x}}_2)$$

Unfortunately the distribution of $g(\underline{x})-k$ is very complex and the problem of finding the exact form is as yet unsolved although it has been studied by many people, including Anderson(1973), Kabe(1963) and Okamoto(1963,1968).

Naturally one can specify the distributions of $g(\underline{x})-k$ given a particular pair of training sets for they will again be normal with means and variances given by,

$$E\{g(\underline{x}) - k \mid w_j\} = (\underline{\mu}_j - \tfrac{1}{2}(\bar{\underline{x}}_1 + \bar{\underline{x}}_2))' \, \underline{S}^{-1} \, (\bar{\underline{x}}_1 - \bar{\underline{x}}_2)$$

$$\mathrm{Var}\{g(\underline{x}) - k \mid w_j\} = (\bar{\underline{x}}_1 - \bar{\underline{x}}_2)' \, \underline{S}^{-1} \, \underline{\Sigma} \, \underline{S}^{-1} \, (\bar{\underline{x}}_1 - \bar{\underline{x}}_2)$$

These distributions however refer specifically to those training sets and any attempt to generalise from them is very dangerous, for example, as we saw in chapter two, they give a falsely optimistic impression of the probability of error.

Alternative methods of estimation of the linear discriminant function for normal data have been suggested in the literature. Notably these include the use of robust estimators and the use of a bias correction , as mentioned in chapter two and returned to in the context of initial misclassification in chapter twelve.

8.3 The robustness of the Bayes classifier

The linear Bayes classifier was derived in the previous section on the basis of four assumptions,

(i) normal distributions

(ii) known covariance matrices

(iii) known prior probabilities

(iv) known paramters

Yet the linear discriminant function has been used in a wide variety of situations and it is important to know just how robust the rule is to departures from these assumptions.

Comparatively few studies have been made of the robustness of the linear classifier to non-normality and those that have been performed have not been conclusive. Lachenbruch, Sneeringer and Revo(1973) simulated data from log-normal, logit normal and arcsinh(normal) distributions and found that the normal based theory did not work well. Whereas Gilbert(1968) and Moore(1973) looked at the performance of the linear discriminant function with discrete data and found that it worked reasonably well. These apparent contradictions are perhaps partly explained by the work on logistic discrimination described in chapter two.

Gilbert(1969) and Marks and Dunn(1974) each studied the importance of having equal covariance matrices. They were only able to conclude that any difference in covariance structure was relatively unimportant if the classes were well separated, and that for small sample sizes the problem of estimating two covariance matrices makes the proper hyperquadratic form even less reliable than the linear approximation.

No studies have been published that look at the estimation of the prior probabilities in isolation, although when training sets are sampled from a mixed population, it is usual to find them estimated by the proportions of each class in the sample.

Lachenbruch(1968) gives a table that serves as a good guide to the situations in which it is reasonable to replace known parameters by maximum likelihood estimates. These results reproduced as table 8.1, show the sample sizes needed to estimate the error rate of the linear discriminant function to within 5% of its true value. The required sample sizes are found to depend upon the degree of separation of the two classes and the dimension of the problem.

Van Ness(1980) considered the performance of the linear discriminant function when the sample sizes fall short of these requirements. The poor quality of the maximum likelihood estimates is found to be sufficient to make the linear discriminant function perform badly in comparison

with other algorithms even though the data are normal.

Table 8.1

Required sample sizes for accurate estimation of the error rate

Dimension	Mahalonbis distance	Sample size
2	1	9
	4	8
	9	7
10	1	35
	4	27
	9	20
20	1	67
	4	51
	9	38

Robustness has been one of the reasons for the use of other estimators of the parameters of the normal distributions. Efron and Morris(1976) and Haff(1980) have, amongst others, considered the use of Stein-like estimators and Peck and Van Ness(1982) looked at other shrinkage estimators, in each case the hope was that the estimators would be more stable in higher dimensions. In all cases improvement was reported but typically it was very small.

8.4 Linear discrimination when covariances differ

Anderson and Bahadur(1962) considered the problem of finding the optimal linear function $g(\underline{x})$ and constant k when the two normally distributed classes have different covariance structures. The results that they derived had been obtained earlier by Kullback(1959) but in a less complete form.

We suppose that the two classes have densities that are p-dimensional normals with mean $\underline{\mu}_j$ and covariance matrices Σ_j, j=1,2. Further let the linear function be,

$$g(\underline{x}) = \underline{a}'\underline{x}$$

so that the values of $\underline{a}'\underline{x}$ are distributed as univariate normals with means $\underline{a}'\underline{\mu}_j$ and variances $\underline{a}'\Sigma_j\underline{a}$. This situation is pictured in figure 8.5 and we are left to find the best values of $\underline{a}$ and k.

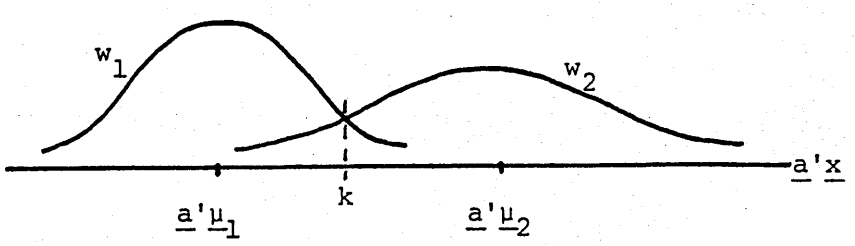


Figure 8.5

Linear discrimination with unequal variances

The probabilities of error under this scheme are clearly

$$1 - \Phi \left| \frac{k - \underline{a}'\underline{\mu}_1}{\sqrt{\underline{a}'\underline{\Sigma}_1 \underline{a}}} \right|$$

and

$$\Phi \left| \frac{k - \underline{a}'\underline{\mu}_2}{\sqrt{\underline{a}'\underline{\Sigma}_2 \underline{a}}} \right|$$

where Φ is the distribution function of the univariate normal. We want to minimise these probabilities, or equivalently, to maximise

$$y_1 = \frac{k - \underline{a}'\underline{\mu}_1}{\sqrt{\underline{a}'\underline{\Sigma}_1 \underline{a}}}$$

and

$$y_2 = \frac{\underline{a}'\underline{\mu}_2 - k}{\sqrt{\underline{a}'\underline{\Sigma}_2 \underline{a}}}$$

Let us first consider the problem of maximising y_1 for a given value of y_2. Eliminating k between the expressions for y_1 and y_2 we obtain,

$$y_1 = \frac{\underline{a}'\underline{d} - y_2 \sqrt{\underline{a}'\underline{\Sigma}_2 \underline{a}}}{\sqrt{\underline{a}'\underline{\Sigma}_1 \underline{a}}}$$

where,

$$\underline{d} = \underline{\mu}_2 - \underline{\mu}_1$$

In order to maximise y_1 for a given y_2 we need to differentiate this expression with respect to $\underline{a}$ giving,

$$\frac{\left| \underline{d} - \frac{y_2 \, \underline{\Sigma}_2 \, \underline{a}}{\sqrt{\underline{a}'\underline{\Sigma}_2 \, \underline{a}}} \right| \sqrt{\underline{a}'\underline{\Sigma}_1 \, \underline{a}} \; - \left| \underline{a}'\underline{d} - y_2 \sqrt{\underline{a}'\underline{\Sigma}_2 \, \underline{a}} \right| \frac{\underline{\Sigma}_1 \, \underline{a}}{\sqrt{\underline{a}'\underline{\Sigma}_1 \, \underline{a}}}}{\underline{a}'\underline{\Sigma}_1 \, \underline{a}}$$

Equating this expression to zero leads to,

$$\underline{d} = (\underline{\Sigma}_1 + \lambda \, \underline{\Sigma}_2) \, \underline{a}$$

where,

$$\lambda = \frac{y_2 \sqrt{\underline{a}'\underline{\Sigma}_2 \, \underline{a}}}{y_1 \sqrt{\underline{a}'\underline{\Sigma}_1 \, \underline{a}}}$$

This equation can be solved iteratively given an initial guess $\underline{a}_o$ and the relationship,

$$\underline{a}_{i+1} = (\underline{\Sigma}_1 + \lambda_i \, \underline{\Sigma}_2)^{-1} \, \underline{d}$$

although to date no conditions for convergence have been published.

Clearly the same approach could be used to maximise y_2 for a given value of y_1. However the minimisation of the total probability of error is more complex, since the usual calculus approach applied to,

$$P(w_1) \; |1 - \Phi(y_1)| + P(w_2) \; |1 - \Phi(y_2)|$$

does not produce equations that can be solved simply and one is forced to use some numerical algorithm.

Just prior to Anderson and Bahadur the same problem had been studied by Cluniés-Ross and Riffenburgh(1960); the difference being that they used a geometric approach. As before the geometry helps one to understand the nature of the problem and we will look at their results, restating them in our notation, but omitting the proofs which may be found in the original paper.

If one has a pair of p-dimensional normal distributions for $\underline{x}$,

$$N(\underline{\mu}_1, \underline{\Sigma}_1) \quad \text{and} \quad N(\underline{\mu}_2, \underline{\Sigma}_2)$$

then it is always possible to find a linear transformation,

$$\underline{y} = \underline{A}\,\underline{x} + \underline{b}$$

such that the distributions become,

$$N(\underline{O}, \underline{I}) \quad \text{and} \quad N(\underline{\mu}, \underline{\Sigma})$$

As well as simplifying the notation this transformation leaves the important characteristics of the configuration unchanges. Thus planes remain planes, tangents remain tangents and the integrals over regions bound by a plane have the same value.The set up for the cases where p=2 is illustrated in figure 8.6.

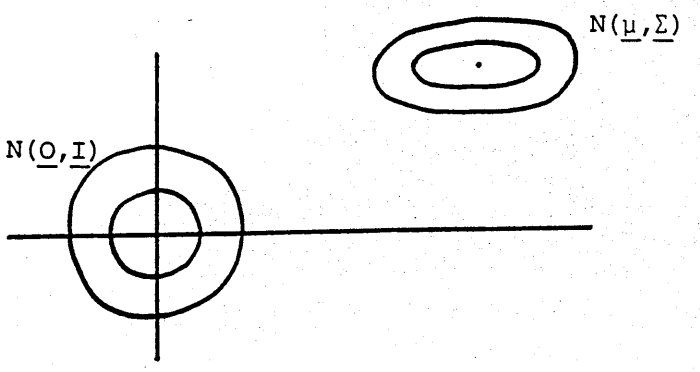


Figure 8.6

Transformed contours in two dimensions

The contours of equal density will have the forms,

$$\underline{y}' \underline{y} = \text{constant}$$

and

$$(\underline{y} - \underline{\mu})' \underline{\Sigma}^{-1} (\underline{y} - \underline{\mu}) = \text{constant}$$

the former being p-dimensional hyperspheres and the later being p-dimensional hyperellipsoids.

The tangents through the point $\underline{y}_o$ will have the equations,

$$(\underline{y} - \underline{\mu})' \underline{\Sigma}^{-1} (\underline{y}_o - \underline{\mu}) = (\underline{y}_o - \underline{\mu})' \underline{\Sigma}^{-1} (\underline{y}_o - \underline{\mu})$$

$$\text{i.e.} \quad (\underline{y} - \underline{y}_o)' \underline{\Sigma}^{-1} (\underline{y}_o - \underline{\mu}) = 0$$

and

$$\underline{y}' \underline{y}_o = \underline{y}_o' \underline{y}_o$$

$$\text{i.e.} \quad (\underline{y} - \underline{y}_o)' \underline{y}_o = 0$$

Clunies-Ross and Riffenburg were able to show that,

'The hyperplane which simultaneously,

(i) effects a fixed division of the probabilities with respect to one multivariate normal population;

(ii) maximises the division of probabilities with respect to another multivariate normal distribution;

is a fully common tangent to the two families of constant likelihood contours'

That is to say the best linear classifier will be a fully common tangent which will thus be a solution of,

$$(\underline{y} - \underline{y}_o)' \underline{y}_o = (\underline{y} - \underline{y}_o)' \underline{\Sigma}^{-1} (\underline{y} - \underline{\mu})$$

This equation allows many solutions for $\underline{y}_o$ the one required being specified by the fixed probability as stated in (i) above.

The problem of solving the fully common tangent equation is completely equivalent to the problem considered earlier in this section. Although this approach does not make it easier to find the best linear classifier what it does do is help our understanding of the problem.

Finally in this section we consider a related problem studied by Guseman, Peters and Walker(1975). They took m general p-dimensional normal distributions and sought the best set of k linear combinations defined by,

$$\underline{y} = \underline{B}\,\underline{x}$$

where $\underline{B}$ is a kxp matrix. The Bayes classifier, based on $\underline{y}$, would be obtained if one found regions S_j such that the probability of correct classification,

$$\sum_{j=1}^{m} \int_{S_j} P(w_j)\, f(\underline{y} \mid w_j)\, d\underline{y}$$

was maximised. Guseman et al thus sought to maximise this quantity both with respect to the regions and the matrix $\underline{B}$.

In fact, although they consider this general problem they were only able to solve the special case where m=2 and k=1. In this case the two distributions transform into univariate normals with unequal variances and Bayes classification is based upon two constants k_1 and k_2 such that

if $k_1 < y < k_2$ classify as w_1

else classify as w_2

They found that even in this restricted case $\underline{B}$, k_1 and k_2 had to be found by numerical optimisation.

8.5 Other Criteria for linear classifier design

In this chapter we consider alternatives to the Bayes criterion for defining a linear classifier. Once again we concentrate on the case where the data are normal.

The linear feature selection methods discussed in chapter five all produce function $g(\underline{x})$ and so could be adapted for classifier design if a method were available for obtaining a suitable value of k.

Like many people since, Fisher(1936) employed a distance measure to select suitable linear features and then turned these features into classifiers by setting k to be midway between the two class means. As is well known this procedure produces the same classifier as the Bayes method with maximum likelihood estimates, see for example, Lachenbruch(1975).

The idea of selecting a feature by maximising the separation was also used by Kullback(1959) except that he based his measure on the concept of information. Any observation $\underline{x}$ is said to carry information relevant to the choice between two classes w_1 and w_2. The information in favour of w_1 and against w_2 is measured by the log

likelihood ratio,

$$I(2:1;\underline{x}) = \ln \left| \frac{f(\underline{x}) | w_1)}{f(\underline{x}) | w_2)} \right|$$

with a similar definition for $I(1:2 ; \underline{x})$.

Given complete knowledge of the distributions it would be natural to define the total information in favour of w_1 as opposed to w_2 obtainable from the whole distributions as,

$$I(2:1) = \int f(\underline{x}|w_1) \ln \left| \frac{f(\underline{x}| w_1)}{f(\underline{x}| w_2)} \right| d\underline{x}$$

Similarly we have,

$$I(1:2) = \int f(\underline{x} |w_2) \ln \left| \frac{f(\underline{x}| w_2)}{f(\underline{x}| w_1)} \right| d\underline{x}$$

and the total information or divergence

$$J(1,2) = I(1:2) + I(2:1)$$

which is a measure of separation between the two classes.

When the two distributions are multivariate normal $N(\underline{\mu}_i, \underline{\Sigma}_i)$ and $\underline{d} = \underline{\mu}_1 - \underline{\mu}_2$, then we find that,

$$I(1:2) = \tfrac{1}{2} \ln \frac{|\underline{\Sigma}_2|}{|\underline{\Sigma}_1|} + \tfrac{1}{2} \operatorname{tr} \left| \underline{\Sigma}_1 (\underline{\Sigma}_2^{-1} - \underline{\Sigma}_1^{-1}) \right| + \tfrac{1}{2} \underline{d}' \underline{\Sigma}_2 \underline{d}$$

with a similar expression for I(2:1) so that,

$$J(1,2) = \tfrac{1}{2}\mathrm{tr}\left|(\underline{\Sigma}_1 - \underline{\Sigma}_2)(\underline{\Sigma}_1^{-1} - \underline{\Sigma}_2^{-1})\right| + \tfrac{1}{2}\underline{d}'\,\underline{\Sigma}_1\,\underline{d} + \tfrac{1}{2}\underline{d}'\underline{\Sigma}_2\,\underline{d}$$

Interestingly when the covariance matrices are equal the divergence reduces to the Mahalanobis distance $\underline{d}'\,\underline{\Sigma}\,\underline{d}$ and the information is equally divided between the two classes with

$$I(1:2) = I(2:1) = \tfrac{1}{2}\,\underline{d}'\,\underline{\Sigma}\,\underline{d}$$

Kullback considered the information and divergence as possible criteria for deriving the best linear discriminant function. Suppose that our function has the form,

$$g(\underline{x}) = \underline{a}'\underline{x}$$

then the information available in the linear combination is measured by,

$$I(1:2) = \tfrac{1}{2}\ln\frac{\underline{a}'\,\underline{\Sigma}_2\,\underline{a}}{\underline{a}'\,\underline{\Sigma}_1\,\underline{a}} - \tfrac{1}{2} + \tfrac{1}{2}\,\frac{\underline{a}'\,\underline{\Sigma}_1\,\underline{a}}{\underline{a}'\,\underline{\Sigma}_2\,\underline{a}} + \tfrac{1}{2}\,\frac{\underline{a}'\,\underline{d}\,\underline{d}'\,\underline{a}}{\underline{a}'\,\underline{\Sigma}_2\,\underline{a}}$$

$$I(2:1) = \tfrac{1}{2}\ln\frac{\underline{a}'\,\underline{\Sigma}_1\,\underline{a}}{\underline{a}'\,\underline{\Sigma}_2\,\underline{a}} - \tfrac{1}{2} + \tfrac{1}{2}\,\frac{\underline{a}'\,\underline{\Sigma}_2\,\underline{a}}{\underline{a}'\,\underline{\Sigma}_1\,\underline{a}} + \tfrac{1}{2}\,\frac{\underline{a}'\,\underline{d}\,\underline{d}'\,\underline{a}}{\underline{a}'\,\underline{\Sigma}_1\,\underline{a}}$$

Whilst when maximising J(1,2),

$$\lambda = \frac{\underline{a}' \Sigma_1 \underline{a} \left| (\underline{a}' \Sigma_2 \underline{a})^2 - (\underline{a}' \Sigma_1 \underline{a})^2 - (\underline{a}' \underline{d})^2 (\underline{a}' \Sigma_1 \underline{a}) \right|}{\underline{a}' \Sigma_2 \underline{a} \left| (\underline{a}' \Sigma_2 \underline{a})^2 - (\underline{a}' \Sigma_1 \underline{a})^2 - (\underline{a}' \underline{d})^2 (\underline{a}' \Sigma_2 \underline{a}) \right|}$$

and

$$\gamma = \frac{\underline{a}' \underline{d} (\underline{a}' \Sigma_1 \underline{a}) \left| \underline{a}' \Sigma_1 \underline{a} + \underline{a}' \Sigma_2 \underline{a} \right|}{(\underline{a}' \Sigma_2 \underline{a})^2 - (\underline{a}' \Sigma_1 \underline{a})^2 + (\underline{a}' \underline{d})^2 (\underline{a}' \Sigma_2 \underline{a})}$$

These equations will be seen to have a general form that is very similar to that derived in section 8.4 and one may try to solve them by a similar iterative process. Kullback(1959) does make some brief comparisons between information based classifiers and estimated Bayes classifiers but bases the comparison on the amount of information in the solution rather than the more usual probability of error, with the result that his findings are hard to generalise.

and

$$J(1,2) = \tfrac{1}{2}\frac{\underline{a}'\,\underline{\Sigma}_2\,\underline{a}}{\underline{a}'\,\underline{\Sigma}_1\,\underline{a}} + \tfrac{1}{2}\frac{\underline{a}'\,\underline{\Sigma}_1\,\underline{a}}{\underline{a}'\,\underline{\Sigma}_2\,\underline{a}} - 1$$

$$+ \tfrac{1}{2}\left|\frac{1}{\underline{a}'\,\underline{\Sigma}_1\,\underline{a}} + \frac{1}{\underline{a}'\,\underline{\Sigma}_2\,\underline{a}}\right| \underline{a}'\,\underline{d}\,\underline{d}'\,\underline{a}$$

Each of these formulae being obtained by substituting the transformed means $\underline{a}'\,\underline{\mu}_i$ and variances $\underline{a}'\,\underline{\Sigma}_i\,\underline{a}$ into the appropriate expressions.

The usual calculus will enable one to derive the equations that need to be solved if these expressions are to be maximised. On differentiating with respect to $\underline{a}$ and equating to zero, one obtains,

$$(\underline{\Sigma}_1 + \lambda\underline{\Sigma}_2)\,\underline{a} = \gamma\underline{d}$$

where when maximising I(1:2)

$$\lambda = \frac{\underline{a}'\,\underline{\Sigma}_1\,\underline{a}}{\underline{a}'\,\underline{\Sigma}_2\,\underline{a}}\left|1 - \frac{(\underline{a}'\,\underline{d})^2}{\underline{a}'\,\underline{\Sigma}_2\,\underline{a} - \underline{a}'\,\underline{\Sigma}_1\,\underline{a}}\right|$$

and

$$\gamma = \frac{(\underline{a}'\,\underline{d})(\underline{a}'\,\underline{\Sigma}_1\,\underline{a})}{(\underline{a}'\,\underline{\Sigma}_2\,\underline{a}) - (\underline{a}'\,\underline{\Sigma}_1\,\underline{a})}$$

8.6 Numerical Problems

We have seen how quite different criteria for defining a linear classifier for two normal distributions lead to the same form of equation, namely,

$$(\underline{\Sigma}_1 - \lambda \underline{\Sigma}_2) \underline{a} = \gamma \underline{d}$$

In fact, Peterson and Mattson(1966) show that if one takes two multivariate normal populations and a linear classifier $\underline{a}'\underline{x}-k$, so that the transformed univariate distributions have means,

$$m_i = \underline{a}' \underline{\mu}_i - k$$

and variances,

$$\sigma_i^2 = \underline{a}' \underline{\Sigma}_i \underline{a} \qquad i = 1,2$$

then any criterion that is a function of these parameters, that is,

$$f(m_1, m_2, \sigma_2^2, \sigma_2^2)$$

will be optimised with respect to $\underline{a}$ when

$$2 \left| \frac{\partial f}{\partial \sigma_1^2} \underline{\Sigma}_1 + \frac{\partial f}{\partial \sigma_1^2} \underline{\Sigma}_2 \right| = (\underline{\mu}_1 - \underline{\mu}_2) \frac{\partial f}{\partial m_2}$$

and

$$\frac{\partial f}{\partial m_1} + \frac{\partial f}{\partial m_2} = 0$$

Thus the Bayes, Fisher and information criteria all lead to the same type of numerical problem.

Kullback(1959) had suggest that an iterative procedure could be used to solve these equations but experiment has shown that convergence is not guaranteed.

Peters and Mattson(1966) recommend a search procedure. Changing their notation slightly we see that we must solve an equation of the form

$$(b_1 \underline{\Sigma}_1 + b_2 \underline{\Sigma}_2) \underline{a} = \underline{d}$$

where b_1 and b_2 are scalars that depend on $\underline{a}$. This equation could be rewritten as,

$$(\xi \cos\theta \underline{\Sigma}_1 = \xi \sin\theta \underline{\Sigma}_2) \underline{a} = d$$

The constant ξ is then seen to be no more than a scaling factor for $\underline{a}$ which will not alter the classifier. Consequently one only needs to consider different angles θ in the range $[-\pi,\pi)$. They suggest that values of θ throughout this range be tried and that these values be used to find

$$\underline{a}_\theta = (\cos\theta\ \underline{\Sigma}_1 + \sin\theta\ \underline{\Sigma}_2)^{-1}\ \underline{d}$$

from which one may find the means and variances of the transformed distributions and hence the value of the criterion. Plotting the criterion against the angles chosen will show the location of the global maximum which can then be located more accurately. The examples produced in that paper demonstate clearly that care needs to be taken with the solution to avoid local optima.

9. PARTIAL DISCRIMINATION

9.1 Introduction

In this chapter we extend the review of linear classification to cover the situation in which some of the cases may be left unclassified. This extension is investigated because it marries naturally with the method of non-parametric feature selection described in chapter six. Under that scheme the 'classified cases' are removed from the training sets and selection continues only with the unclassified cases.

It is common in classification to find that some of the cases fall between two classes in such a way that their origin is doubtful. It is sensible to allow such cases to remain unclassified. This is the reject option briefly reviewed by Hand(1981), although it is perhaps more usually referred to as partial discrimination .

Despite the appeal of the idea it has rarely been used. In this chapter we will first describe the small amount of work that has been done and then consider the problem of defining a partial classifier. The numerical problems associated with the partial classification of normal data are solved and the algorithm is applied to an example taken from the literature.

9.2 Previous work on partial discrimination

The idea underlying partial discrimination has been recognised for many years, for example, it is mentioned but not pursued in Kendall and Stuart(1966). It seems always to have been primarily linked with non-parametric procedures and it is in this context that Quesenberry and Gessaman(1968) published a partial discrimination scheme based on tolerance regions. They suggested the use of two such regions A_1 and A_2 together with the decision rule,

if $\underline{x} \in A_1$ classify $\underline{x}$ as w_1

if $\underline{x} \in A_2$ classify $\underline{x}$ as w_2

if $\underline{x} \in (A_1 \cap A_2) \cup (A_1' \cap A_2')$ reserved judgement

Although their concern was primarily with non-parametric discrimination they did quote one theorem relevant to the definition of a parametric partial discrimination scheme. They showed, by a simple extension of the results of Welch(1939), that in order to form a partial discrimination scheme with fixed probabilities,

$$P(\underline{x} \in A_1 \mid w_1) \quad \text{and} \quad P(\underline{x} \in A_2 \mid w_2)$$

then one should use regions defined in terms of the likelihood ratios. That is,

$$A_1 = \{ \underline{x} ; \frac{f(\underline{x} \mid w_1)}{f(\underline{x} \mid w_2)} > c_1 \}$$

and

$$A_2 = \{ \underline{x} ; \frac{f(\underline{x} \mid w_2)}{f(\underline{x} \mid w_1)} > c_2 \}$$

This result is expanded upon in the next section.

Broffitt, Randles and Hogg(1976) also looked at non-parametric discrimination using a partial classifier based on ranks. They supposed that they had training sets of size n_1 and n_2 together with a classification function $g(\underline{x})$, which we will suppose tends to give smaller values when the case is from class w_1.

The values from training set one may be transformed according to $g(\underline{x})$ and the resulting values ranked producing,

$$R_i = \text{Rank}(g(\underline{x}_i))$$

In which cases $(R_1,\ldots,R_{n_1})$ will be uniformly distributed over the set of integers $1,2,\ldots,n_1$. Using this idea they suggest that the case of unknown origin, $\underline{x}$, be ranked along with each of the training sets in turn. Selecting what seem suitable vaules for a_1 and a_2 the scheme is defined in terms

of

$$A_1 = \{ \underline{x} ; \text{rank}_1 (g(\underline{x})) < a_1 (n_1 + 1) \}$$

$$A_2 = \{ \underline{x} ; \text{rank}_2 (g(\underline{x})) > a_2 (n_2 + 1) \}$$

Although this procedure could be used with any function $g(\underline{x})$, they only look at the use of Fisher's linear discriminant function.

This scheme is contrasted by means of simulation experiments with a tolerance region approach along the lines of Quesenberry and Gessaman(1966) which used the actual order statistics of $g(\underline{x}_i)$, and a third normal based method. This later method, illustrated in figure 9.1, used regions defined by,

$$A_1 = \{ \underline{x} ; g(\underline{x}) < \mu_1 + \alpha \sigma_1 \}$$

$$A_2 = \{ \underline{x} ; g(\underline{x}) > \mu_2 - \alpha \sigma_2 \}$$

Neither in this paper nor in a later one, Randles, Broffitt, Ramberg and Hogg(1978), which used a quadratic function for $g(\underline{x})$, where they able to find a universally best scheme, although they do recommend the use of their own rank based procedure.

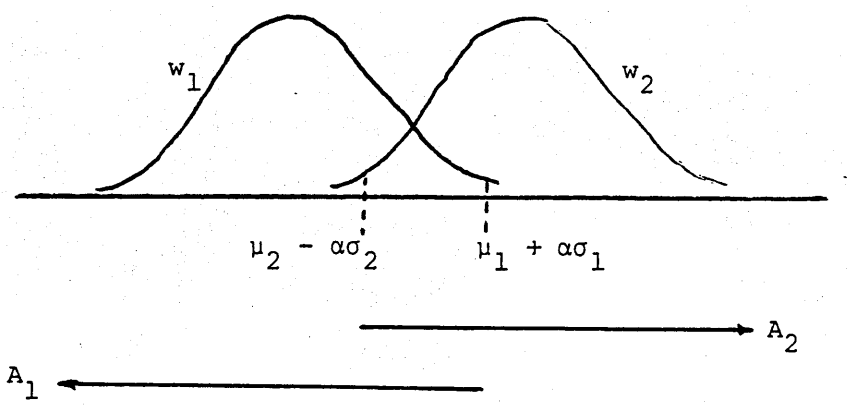


Figure 9.1

Partial discrimination using tolerance intervals

9.3 Defining partial classifiers

As we have seen a total classification scheme may be thought of as a rule that divides the measurement space S of $\underline{x}$ into mutually exclusive and exhaustive regions S_j within which cases are assigned to class w_j. Using the usual notation for the densities and prior probabilities and letting c_{ij} denote the cost of incorrectly classifying a case from class w_j as coming from class w_i, we might choose S_j so as to minimise the total cost of misclassification,

$$\sum_{j=1}^{m} \sum_{\substack{i=1 \\ i \neq j}}^{m} \int_{S_j} c_{ij} \, f(\underline{x} \mid w_j) \, P(w_j) \, d\underline{x}$$

This is the usual Bayes classification.

For a partial classification scheme the only difference is that the regions S_j are no longer exhaustive, in that,

$$\bigcup_j S_j \neq S$$

Any points that lie within

$$S_* = \{ \bigcup_j S_j \}'$$

are to be denoted as unclassified and we may associate a cost c_j with any unclassified case that actually comes from

class w_j.

The equivalent to Bayes classification would now be achieved if one chooses S_j so as to minimise,

$$\sum_{j=1}^{m} \sum_{\substack{i=1 \\ i\neq j}}^{m} \int_{S_j} c_{ij}\, f(\underline{x} \mid w_j) P(w_j)\, d\underline{x} + \sum_{j=1}^{m} \int_{S_*} c_j\, f(\underline{x} \mid w_j) P(w_j)\, d\underline{x}$$

This general optimisation problem is very complex but as we will see it has many similarities with total classification, and corresponding simplifying assumptions may be made.

Suppose that one either has a single measurement or that one selects a single feature, $g(\underline{x})$, so that the two possible classes may be represented as in figure 9.2.

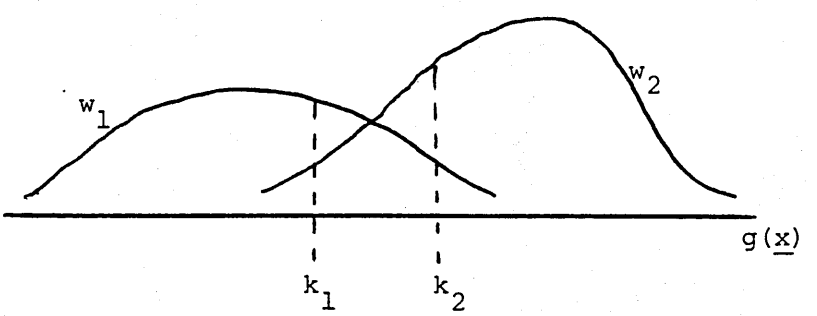


Figure 9.2

Two class partial discrimination

Partial discrimination would now be achieved if one were to select values k_1 and k_2 such that,

if $g(\underline{x}) < k_1$ classify as w_1

if $g(\underline{x}) > k_2$ classify as w_2

otherwise unclassified

One clearly cannot guarantee that this will work sensibly in all situations, for the densities might resemble those in figure 9.3. However as with total classification, the method will usually work in practice.

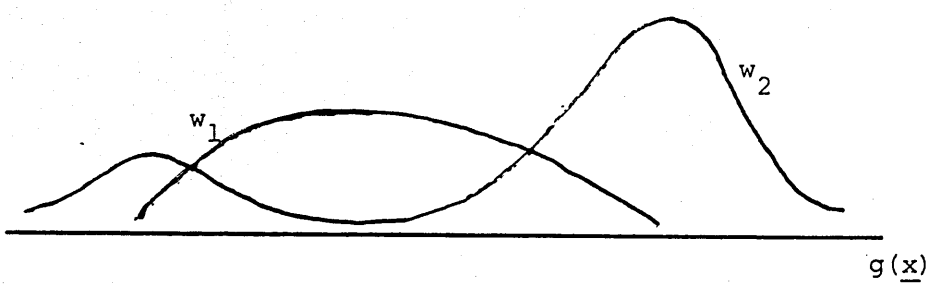


Figure 9.3

Two populations unsuitable for partial discrimination

The problem is then nearly identical to that considered in chapter eight. We need firstly to define the function $g(\underline{x})$ and then to select the threshold values k_1 and k_2. These choices could be based on any one of a number of criteria, for example,

(i) to give the minimum probability (cost) of error for a specified probability (cost) of an unclassified case.

(ii) to give the minimum probability (cost) of an unclassified case for a specified probability (cost) of error.

(iii)to minimise the total cost

Suppose firstly that one were to follow criterion (i) above. This would require us to minimise the cost of error,

$$c_{21} \int_{S_1} f(\underline{x} \mid w_1)P(w_1)\ d\underline{x} + c_{12} \int_{S_2} f(\underline{x} \mid w_2)P(w_2)\ d\underline{x}$$

subject to a fixed cost C associated with unclassified cases, where,

$$C = c_1 \int_{S_*} f(\underline{x} \mid w_1)P(w_1)\ d\underline{x} + c_2 \int_{S_*} f(\underline{x} \mid w_2)P(w_2)\ d\underline{x}$$

Having selected a feature $t=g(\underline{x})$ one would be attempting to discriminate using the two univariate densities, $f(t \mid w_1)$

and $f(t \mid w_2)$. The values for k_1 and k_2 would be chosen so as to minimise,

$$c_{21} \int_{k_2}^{\infty} f(t \mid w_1)P(w_1)\, dt \;+\; c_{12} \int_{-\infty}^{k_1} f(t \mid w_2)P(w_2)\, dt$$

subject to,

$$C \;=\; c_1 \int_{k_1}^{k_2} f(t \mid w_1)P(w_1)\, dt \;+\; c_2 \int_{k_1}^{k_2} f(t \mid w_2)P(w_2)\, dt$$

This optimisation may be achieved by the use of Lagrange multipliers, leading us to differentiate

$$c_{21} \int_{k_2}^{\infty} f(t \mid w_1)P(w_1)\, dt \;+\; c_{12} \int_{-\infty}^{k_1} f(t \mid w_2)P(w_2)\, dt$$

$$+\; \lambda \left[C - c_1 \int_{k_1}^{k_2} f(t \mid w_1)P(w_1)\, dt \;-\; c_2 \int_{k_1}^{k_2} f(t \mid w_2)P(w_2)\, dt\right]$$

with respect to k_1 and k_2, giving when equated to zero,

$$c_{12}\, f(k_1 \mid w_2)P(w_2) + \lambda c_1\, f(k_1 \mid w_1)P(w_1) + \lambda c_2\, f(k_1 \mid w_2)P(w_2) \;=\; 0$$

and

$$c_{21}\, f(k_2 \mid w_1)P(w_1) + \lambda c_1\, f(k_2 \mid w_1)P(w_1) + \lambda c_2\, f(k_2 \mid w_2)P(w_2) \;=\; 0$$

Eliminating λ we obtain,

$$\lambda = \frac{c_{12}\, f(k_1 \mid w_2)P(w_2)}{c_1\, f(k_1 \mid w_1)P(w_1) + c_2\, f(k_1 \mid w_2)P(w_2)}$$

Which when rearranged gives,

$$\frac{f(k_1 \mid w_1)}{f(k_1 \mid w_2)} = \frac{c_{12} - \lambda c_2}{-\lambda c_1} \frac{P(w_2)}{P(w_1)}$$

$$\frac{f(k_2 \mid w_2)}{f(k_2 \mid w_1)} = \frac{c_{21} - \lambda c_1}{-\lambda c_2} \frac{P(w_1)}{P(w_2)}$$

The significance of the general result becomes more apparent when one considers the special case whereby $c_{12}=c_{21}$ and $c_1=c_2$ for then,

$$\frac{f(k_1 \mid w_1)P(w_1)}{f(k_1 \mid w_2)P(w_2)} = \frac{f(k_2 \mid w_2)P(w_2)}{f(k_2 \mid w_1)P(w_1)}$$

and if $P(w_1)=P(w_2)$ then,

$$\frac{f(k_1 \mid w_1)}{f(k_1 \mid w_2)} = \frac{f(k_2 \mid w_2)}{f(k_2 \mid w_1)}$$

Thus, in these special cases, we need to choose k_1 and k_2 so that their likelihood ratios are equal.

If one were to choose the second of the criteria suggested above then one would need to minimise the cost of the unclassified cases,

$$c_1 \int_{k_1}^{k_2} f(t \mid w_1)\, P(w_1)\, dt \;+\; c_2 \int_{k_1}^{k_2} f(t \mid w_2)\, P(w_2)\, dt$$

for a specified cost of error, C, where,

$$C \;=\; c_{21} \int_{k_2}^{\infty} f(t \mid w_1)\, P(w_1)\, dt \;+\; c_{12} \int_{-\infty}^{k_1} f(t \mid w_2)\, P(w_2)\, dt$$

Once again this problem may be solved by the use of Lagrange multipliers, leading us to consider the function,

$$c_1 \int_{k_1}^{k_2} f(t \mid w_1)\, P(w_1)\, dt \;+\; c_2 \int_{k_1}^{k_2} f(t \mid w_2)\, P(w_2)\, dt$$

$$+\; \lambda \left| C - c_{21} \int_{k_2}^{\infty} f(t \mid w_1)\, P(w_1)\, dt \;-\; c_{12} \int_{-\infty}^{k_1} f(t \mid w_2)\, P(w_2)\, dt \right|$$

Which on differentiating with respect to k_1 and k_2 produces equations,

$$c_1\, f(k_1 \mid w_1) P(w_1) + c_2\, f(k_1 \mid w_2)\, P(w_2) + \lambda c_{12}\, f(k_1 \mid w_2) P(w_2) \;=\; 0$$

$$c_1\, f(k_2 \mid w_1) P(w_1) + c_2\, f(k_2 \mid w_2)\, P(w_2) + \lambda c_{21}\, f(k_2 \mid w_1) P(w_1) \;=\; 0$$

so that,

$$\frac{f(k_1 \mid w_1)}{f(k_1 \mid w_2)} = \frac{c_2 + \lambda c_{12}}{-c_1}$$

$$\frac{f(k_2 \mid w_2)}{f(k_2 \mid w_1)} = \frac{c_1 + \lambda c_{21}}{-c_1}$$

If λ is eliminated from this pair of equations one finds that,

$$\frac{c_{12}\, f(k_1 \mid w_2)\, P(w_2)}{c_1\, f(k_1 \mid w_1)\, P(w_1) + c_2\, f(k_1 \mid w_2)\, P(w_2)}$$

$$= \frac{c_{21}\, f(k_2 \mid w_1)\, P(w_1)}{c_1\, f(k_2 \mid w_1)\, P(w_1) + c_2\, f(k_2 \mid w_2)\, P(w_2)}$$

which is exactly the form of the relationship derived for the first criterion.

Finally we might choose the third of our suggested criteria, that is to minimise the total cost,

$$c_{21} \int_{k_2}^{\infty} f(t \mid w_1)\, P(w_1)\, dt + c_{12} \int_{-\infty}^{k_1} f(t \mid w_2)\, P(w_2)\, dt$$

$$+ c_1 \int_{k_1}^{k_2} f(t \mid w_1)\, P(w_1)\, dt + c_2 \int_{k_1}^{k_2} f(t \mid w_2)\, P(w_2)\, dt$$

If we differentiate with respect to k_1 and k_2 and equate to zero we obtain,

$$c_{12}\, f(k_1 \mid w_2)\, P(w_2) - c_1\, f(k_1 \mid w_1)\, P(w_1) - c_2\, f(k_1 \mid w_2)\, P(w_2) = 0$$

and

$$c_{21}\, f(k_2 \mid w_1)\, P(w_1) - c_1\, f(k_2 \mid w_1)\, P(w_1) - c_2\, f(k_2 \mid w_2)\, P(w_2) = 0$$

Thus to find k_1 we need,

$$\frac{f(k_1 \mid w_1)}{f(k_1 \mid w_2)} = \frac{c_{12} - c_2}{c_1}\,\frac{P(w_2)}{P(w_1)}$$

and

$$\frac{f(k_2 \mid w_2)}{f(k_2 \mid w_1)} = \frac{c_{21} - c_1}{c_2}\,\frac{P(w_1)}{P(w_2)}$$

It is not surprising that these equations have no solution when $c_{12}=c_2$ and $c_{21}=c_1$ for if these costs are equal we are not distinguishing between errors and unclassified cases. In practice it must be that unclassified cases are less costly than errors, i.e.,

$$c_1 < c_{21}$$

$$c_2 < c_{12}$$

9.4 Two Normal Distributions

We will now consider the problem of finding the linear classifier that minimises the total cost of partial discrimination when the two classes are multivariate normal but with differing covariance structures. Suppose that the two normal distributions are $N(\underline{\mu}_1 , \underline{\Sigma}_1)$ and $N(\underline{\mu}_2 , \underline{\Sigma}_2)$. The linear classifier will have the form,

$$g(\underline{x}) = \underline{a}' \underline{x}$$

and we will suppose that the classification rule is of the form,

$\underline{a}' \underline{x} < k_1$ classify as w_1

$\underline{a}' \underline{x} > k_2$ classify as w_2

else unclassified

This is the situation illustrated in figure 9.2 except that we now know that the univariate distributions are normal with means $m_1 = \underline{a}' \underline{\mu}_1$ and $m_2 = \underline{a}' \underline{\mu}_2$ and variances $\sigma_i^2 = \underline{a}' \underline{\Sigma}_i \underline{a}$

Using Φ to represent the cumulative normal distribution we may express the total cost R as,

$$R = c_{21}\{1 - \Phi(\frac{k_2 - m_1}{\sigma_1})\} P(w_1) + c_{21}\Phi(\frac{k_1 - m_2}{\sigma_2}) P(w_2)$$

$$+ c_1\{\Phi(\frac{k_2 - m_1}{\sigma_1}) - \Phi(\frac{k_1 - m_1}{\sigma_1})\} P(w_1)$$

$$+ c_2\{\Phi(\frac{k_2 - m_2}{\sigma}) - \Phi(\frac{k_1 - m_2}{\sigma_2})\} P(w_2)$$

Using the results from the previous section we know that,

$$\frac{P(w_1)c_1}{\sigma_1} \phi\left\{\frac{k_1 - m_1}{\sigma_1}\right\} = \frac{(c_{12} - c_2) P(w_2)}{\sigma_2} \phi\left\{\frac{(k_1 - m_2)}{\sigma_2}\right\}$$

and

$$\frac{P(w_2)c_2}{\sigma_2} \phi\left\{\frac{k_2 - m_2}{\sigma_2}\right\} = \frac{(c_{21} - c_1) P(w_1)}{\sigma_1} \phi\left\{\frac{(k_2 - m_1)}{\sigma_1}\right\}$$

where ϕ represents the univariate normal density.

Thus given a linear classifier $a'x$ we may deduce m_i and σ_i and hence we may solve for k_1 and k_2. These equations are in fact quadratics in k_i and it is important to chack that one really does have the solution that minimises R.

In order to minimise R with respect to a we need to differentiate. Using,

$$\frac{\partial R}{\partial \underline{a}} = \frac{\partial R}{\partial m_1}\frac{\partial m_1}{\partial \underline{a}} + \frac{\partial R}{\partial m_2}\frac{\partial m_2}{\partial \underline{a}} + \frac{\partial R}{\partial \sigma_1^2}\frac{\partial \sigma_1^2}{\partial \underline{a}} + \frac{\partial R}{\partial \sigma_2^2}\frac{\partial \sigma_2^2}{\partial \underline{a}}$$

so that at the minimum,

$$\frac{\partial R}{\partial \underline{a}} = 0$$

Now,

$$\frac{\partial m_1}{\partial \underline{a}} = \underline{\mu}_1 \quad \frac{\partial m_2}{\partial \underline{a}} = \underline{\mu}_2 \quad \frac{\partial \sigma_1^2}{\partial \underline{a}} = 2\underline{\Sigma}_1 \underline{a} \quad \frac{\partial \sigma_2^2}{\partial \underline{a}} = 2\underline{\Sigma}_2 \underline{a}$$

so that at the minimum

$$-\left\{ \frac{\partial R}{\partial m_1} \underline{\mu}_1 + \frac{\partial R}{\partial m_2} \underline{\mu}_2 \right\} = 2 \left\{ \frac{\partial R}{\partial \sigma_1^2} \underline{\Sigma}_1 + \frac{\partial R}{\partial \sigma_2^2} \underline{\Sigma}_2 \right\} \underline{a}$$

with

$$\frac{\partial R}{\partial m_1} = \frac{c_{21} P(w_1)}{\sigma_1} \phi \left\{ \frac{k_2 - m_1}{\sigma_1} \right\} - \frac{c_1 P(w_1)}{\sigma_1} \phi \left\{ \frac{k_2 - m_1}{\sigma_1} \right\} + \frac{c_1 P(w_1)}{\sigma_1} \phi \left\{ \frac{k_1 - m_1}{\sigma_1} \right\}$$

$$\frac{\partial R}{\partial m_2} = \frac{c_{12} P(w_2)}{\sigma_2} \phi \left\{ \frac{k_1 - m_2}{\sigma_2} \right\} - \frac{c_2 P(w_2)}{\sigma_2} \phi \left\{ \frac{k_2 - m_2}{\sigma_2} \right\} + \frac{c_2 P(w_2)}{\sigma_2} \phi \left\{ \frac{k_1 - m_2}{\sigma_2} \right\}$$

$$\frac{\partial R}{\partial \sigma_1^2} = \frac{c_{21} P(w_1)}{\sigma_1^3} \; \phi \left\{ \frac{k_2 - m_1}{\sigma_1} \right\} (k_2 - m_1)$$

$$- \frac{c_1 P(w_1)}{\sigma_1^3} \; \phi \left\{ \frac{k_2 - m_1}{\sigma_1} \right\} (k_2 - m_1)$$

$$+ \frac{c_1 P(w_1)}{\sigma_1^3} \; \phi \left\{ \frac{k_1 - m_1}{\sigma_1} \right\} (k_1 - m_1)$$

$$\frac{\partial R}{\partial \sigma_2^2} = \frac{-c_{12} P(w_2)}{\sigma_2^3} \; \phi \left\{ \frac{k_1 - m_2}{\sigma_2} \right\} (k_1 - m_2)$$

$$- \frac{c_2 P(w_2)}{\sigma_2^3} \; \phi \left\{ \frac{k_2 - m_2}{\sigma_2} \right\} (k_2 - m_2)$$

$$+ \frac{c_2 P(w_2)}{\sigma_2^3} \; \phi \left\{ \frac{k_1 - m_2}{\sigma_2} \right\} (k_1 - m_2)$$

The procedure advocated by Peterson and Mattson(1966) is no longer available to us since all four of the derivatives depend upon the threshold values k_i. However we may still attempt either to optimise directly or to try an iterative process, such as,

(a) make an initial guess a_o

(b) derive m_i and o_i and hence k_1 and k_2

(c) Evaluate the partial derivatives

(d) Solve

$$a_{i+1} = -\tfrac{1}{2} \left| \frac{\partial R}{\partial \sigma_1^2} \underline{\Sigma}_1 + \frac{\partial R}{\partial \sigma_2^2} \underline{\Sigma}_2 \right|^{-1} \left| \frac{\partial R}{\partial m_1} \underline{\mu}_1 + \frac{\partial R}{\partial m_1} \underline{\mu}_2 \right|$$

to obtain an updated estimate

(e) if convergence is not obtained return to (b)

To illustate this algorithm the example from Peterson and Mattson has been used. Here we have two bivariate normal distributions with parameters,

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \begin{pmatrix} 4 & 1 \\ 1 & 1 \end{pmatrix} \qquad P(w_1)=0.4$$

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad \begin{pmatrix} 2 & 1 \\ 1 & 4 \end{pmatrix} \qquad P(w_2)=0.6$$

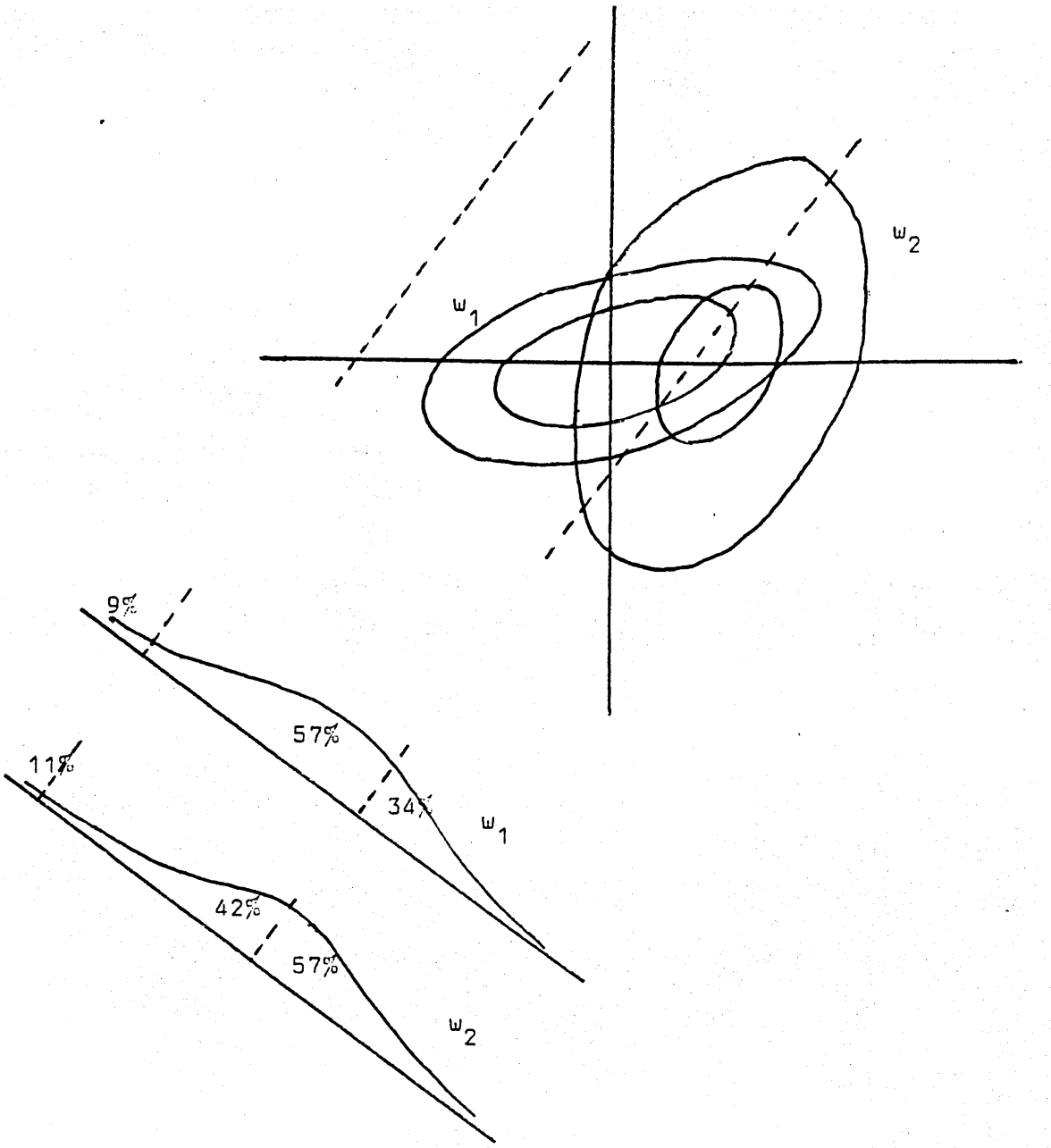


Figure 9.4

The partial discrimination solution to

Peterson & Mattson's problem

10 A MEASURE OF SEPARATION FOR USE IN DEFINING NON-PARAMETRIC DISCRIMINATION SCHEMES

10.1 Introduction

Non-parametric discrimination has a long history in pattern recognition and has usually involved one of three approaches; either non-parametric density estimation, nearest neighbour analysis or more rarely the use of tolerance intervals. As we saw in chapter nine it is this later approach, introduced by Quesenberry and Gessman(1968) that links naturally with partial discrimination.

In this chapter we will firstly consider the definitions basic to tolerance intervals and then consider the characteristics of a non-parametric measure of separation that can be used to define a discrimination scheme. As with any other measure of separation we can define the scheme by looking for the linear function of the original variables that maximises that measure. The scheme is very closely related to the method of non-parametric feature selection discussed in chapter six and many of the results presented here are also relevant to that earlier work. The method is used to analyse the ballistocardiograms and psychiatric data in chapter eleven.

10.2 Tolerance Intervals

This section contains only the briefest of introductions to tolerance intervals. For a fuller account with proofs see Guttman(1970).

Suppose that X is a random variable with continous density f(x) defined for all real x. If we collect a sample of data and use it to estimate two values,

$$l_1(x_1, x_2, \ldots, x_n) \quad \text{and} \quad l_2(x_1, x_2, \ldots, x_n)$$

then the integral,

$$C = \int_{l_1}^{l_2} f(x)\,dx$$

is known as the coverage of the interval (l_1, l_2) and the interval is a b-content region with confidence γ if,

$$P(C \geq b) \geq \gamma$$

The problem, having chosen b and γ, is to find a method for estimating l_1 and l_2 such that the definition is satisfied.

Generally the best estimates for l_1 and l_2 will depend upon the form of the density f(x), but Wilks(1941) showed that if we take the r^{th} order statistic $x_{(r)}$ and the

$(n-r+1)^{th}$ order statistic $x_{(n-r+1)}$ as the values for l_1 and l_2 respectively then the resulting interval has a coverage that follows a beta distribution $I_{n-2r+1,2r}$ regardless of the distribution of X. This powerful result enables one to use tables of the beta distribution to set up non-parametric tolerance intervals based on the order statistics, although, of course, narrower intervals may be obtained if the form of the density is known.

Tukey(1947) introduced the term 'blocks' to refer to the (n+1) intervals,

$$(-\infty, x_{(1)}) \quad (x_{(1)}, x_{(2)}) \quad \ldots\ldots \quad (x_{(n)}, \infty)$$

These blocks are said to be statistically equivalent because they all have expected coverages equal to 1/(n+1). In fact the total coverage of any k blocks can be shown to follow a beta distribution $I_{k,n-k+1}$.

The only form of density f(x) that has received any appreciable consideration is the univariate normal. Wilks(1941) showing that the interval,

$$\bar{x} \pm k\,s$$

where $\bar{x}$ and s are the sample mean and standard deviation, has expected coverage b if k is taken to be,

$$\sqrt{\frac{n+1}{n}}\, t_{n-1}\,(1 - \tfrac{1}{2}b)$$

The problem of specifying k for given values of b and γ is more complex. Wald and Wolfowitz(1946) gave an approximate solution whereby,

$$k = k_1 k_2$$

with k_1 the solution of,

$$\Phi\left(\frac{1}{\sqrt{n}} + k_1\right) - \Phi\left(\frac{1}{\sqrt{n}} - k_1\right) = b$$

and

$$k_2 = \sqrt{\frac{n-1}{\chi^2_{n-1,\gamma}}}$$

Others such as Bowker(1947) and Weissberg and Beath(1960) have given tables for finding k.

10.3 A Non-Parametric Measure of Separation

We saw in chapter eight how it is possible to define a linear classifier $g(\underline{x})=\underline{a}'\underline{x}$ by a host of different methods. One of the more popular techniques is to choose $\underline{a}$ so as to maximise some measure of the separation between the classes. In this section we will define a non-parametric measure of separation that is based on tolerance intervals, that is simple to handle and that works well in practice.

The vast majority of measures of separation that have been suggested in the literature assume that the forms of the densities are known and the few that don't tend to rely on the first two sample moments, implicitly assuming aa approximate normal structure.

The proposed measure, which will be denoted by P, would for known distributions be based on the quantiles. Suppose that we have two univariate densities as illustrated in figure 10.1, then if without any loss of generality, we assume that,

$$x_{1(q)} > x_{2(q)}$$

where,

$$F_i(x_i(p)) = p$$

and $p=1-q$, $q>p$ then we may define the separation to be,

$$R_{(p)} = \frac{x_{1(p)} - x_{2(q)}}{x_{1(q)} - x_{2(p)}}$$

It will be seen from the illustrations that R expresses the overlap or separation between the two distributions as a proportion of the combined range.

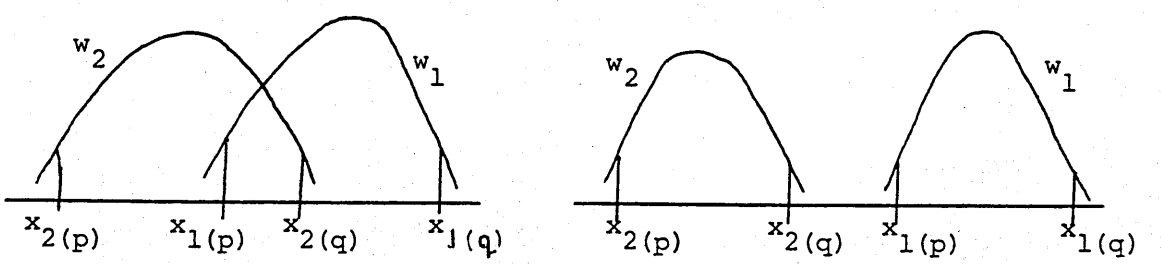


Figure 10.1
Values used in the calculation of $R_{(p)}$

This measure would give poor results in cases such as those illustrated in figure 10.2, where the overlap is total. However, such cases present no real problem as it is always possible to transform to the absolute deviation from the centre of the narrower density so reverting to the situation shown in figure 10.1.

If one takes precautions to guard against complete overlap the R will always be a proportion, negative if the distributions overlap and positive if they are separated.

Thus,

$$-1 \leq R_{(p)} \leq 1$$

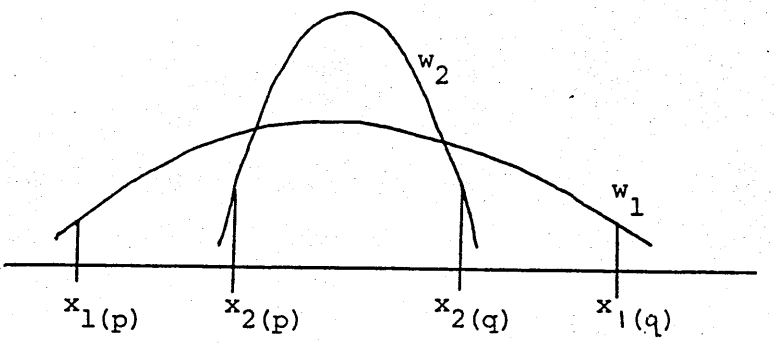


Figure 10.2
A configuration for which $R_{(p)}$ is unsuitable.

As it stands R is dependent on the form of the distributions F, but the work on tolerance regions suggests that we might make a non-parametric version,

$$R_{(rs)} = \frac{x_{1(r)} - x_{2(s)}}{x_{1(s)} - x_{2(r)}}$$

where s>r and $x_{i(r)}$ is the rth order statistic of the sample from population w_i. The order statistics serve as non-parametric estimates of $x_{i(p)}$ since,

$$E(F_i(X_{i(r)}) = \frac{r}{n+1}$$

and r can be chosen so that,

$$\frac{r}{n+1} \approx p$$

Before considering the properties of $R_{(rs)}$ it is instructive to note the relationship between R and an established measure of separation commonly used for normal data.

If we have two normal distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ with known parameters then

$$x_{i(p)} = \mu_i - \alpha\sigma_i$$

$$x_{i(q)} = \mu_i + \alpha\sigma_i$$

where

$$\alpha = \Phi^{-1}(q)$$

Thus,

$$R_{(p)} = \frac{(\mu_1 - \alpha\sigma_1) - (\mu_2 + \alpha\sigma_2)}{(\mu_1 + \alpha\sigma_1) - (\mu_2 - \alpha\sigma_2)}$$

$$= 1 - \frac{2\alpha}{\alpha + \dfrac{(\mu_1 - \mu_2)}{(\sigma_1 + \sigma_2)}}$$

So that R is directly linked to the measure

$$\frac{\mu_1 - \mu_2}{\frac{1}{2}(\sigma_1 + \sigma_2)}$$

Had the observations $\underline{x}$ come from multivariate normal distributions with a common covariance matrix then the linear combination that maximises R would also maximise,

$$\frac{\underline{a}'(\underline{\mu}_1 - \underline{\mu}_2)}{\sqrt{\underline{a}' \, \underline{\Sigma} \, \underline{a}}}$$

that is to say it would be the usual linear discriminant function.

10.4 The Sampling Distribution of $R_{(rs)}$

Generally the sampling distribution is very complex but as we will see it is possible to approximate the mean and standard deviation so enabling a partial summary. We concentrate on the normal case since that is the most likely to be of interest, but the method is general enough to be applied to any suspected underlying structure.

Let

$$u = x_{1(r)} - x_{2(s)}$$

$$v = x_{1(s)} - x_{2(r)}$$

so that

$$R_{(rs)} = \frac{u}{v}$$

Assuming that we have two random samples,

$$E(u) = E(x_{1(r)}) - E(x_{2(s)})$$

$$Var(u) = Var(x_{1(r)}) + Var(x_{2(s)})$$

$$E(v) = E(x_{1(s)}) - E(x_{2(r)})$$

$$Var(v) = Var(x_{1(s)}) + Var(x_{2(r)})$$

$$Cov(u,v) = Cov(x_{1(r)}, x_{1(s)}) + Cov(x_{2(r)}, x_{2(s)})$$

Taking the usual approximations for the moments of a ratio, see for example, Kendall and Stuart(1966),

$$E(R_{(rs)}) = E(u)/E(v)$$

$$Var(R_{(rs)}) = \left|\frac{E(u)}{E(v)}\right|^2 \left|\frac{Var(u)}{E(u)^2} - \frac{2Cov(u,v)}{E(u)E(v)} + \frac{Var(v)}{E(v)^2}\right|$$

Using these formulae we may relate the mean and standard error of $R_{(rs)}$ to the properties of the order statistics of the original data.

The joint distribution of $x_{i(r)}$ and $x_{i(s)}$ can be written in terms of the distribution functions F_i as,

$$\frac{F_i(x_{i(r)})^{r-1}|F_i(x_{i(s)}) - F_i(x_{i(r)})|^{s-r-1}\,|1-F_i(x_{i(s)})|^{n_i-s}dF_i(x_{i(r)})dF_i(x_{i(s)})}{B(r,\ s-r)\ B(s,\ n-s+1)}$$

where n_i are the sample sizes and B is the beta function.

Fortunately it is not necessary that we evaluate the moments of this exact distribution because David and Johnson(1954) derived expansions for the moments using a method due to Karl Pearson whereby the order statistic $x_{i(r)}$ is obtained as a Taylor series about the true value $X_{i(r)}$,

$$F_i(X_{i(r)}) = \frac{r}{n+1} = p_r$$

This method leads to expressions for the moments of $x_{i(r)}$ in terms of $X_{i(r)}$ and its derivatives,

$$x^k_{i(r)} = \frac{d^k X_{i(r)}}{dF_i} \qquad k = 1, 2, \ldots$$

Ignoring terms in $(n+2)^{-3}$ the approximations to the first two moments become, (dropping the suffix i),

$$E(x_{(r)}) = X_r + \frac{p_r q_r}{2(n+2)} X_r^{11} + \frac{p_r q_r}{(n+2)^2}\left|\frac{(q_r - p_r)X_r^{111}}{3} + \frac{pqX_r^{IV}}{8}\right|$$

$$Var(x_{(r)}) = \frac{p_r q_r}{(n+2)} X_r^1 + \frac{p_r q_r}{(n+2)^2}\left|2(q_r-p_r)X_r^1 X_r^{11} + p_r q_r(X_r^1 X_r^{11} + \tfrac{1}{2}X_r^{11^2})\right|$$

$$Cov(x_{(r)},x_{(s)}) = \frac{p_r q_s}{(n+2)} X_r^1 X_s^1 + \frac{p_r q_s}{(n+2)^2}\left|(q_r - p_s) X_r^{11} X_s^{11}\right.$$

$$\left.+(q_s-p_s)X_r^1 X_s^{11} + \tfrac{1}{2}p_r q_r X_r^{111}X_s^1 + \tfrac{1}{2}p_s q_s X_r^1 X_s^{111} + \tfrac{1}{2}p_r q_s X_r^{11}X_s^{11}\right|$$

In the case where the data is normally distributed,

i.e. $X \sim N(\mu,\sigma^2)$

then

$$X_r = \mu + \sigma \Phi^{-1}\left(\frac{r}{n+1}\right)$$

$$X_r^{1} = \sigma\sqrt{2\pi}\ \exp\left[-\tfrac{1}{2}(X_r - \mu)^2/\sigma^2\right]$$

$$X_r^{11} = 2\pi(X_r - \mu)\ \exp\left[(X_r - \mu)^2/\sigma^2\right]$$

$$X_r^{111} = \sigma(2\pi)^{\frac{3}{2}}\ (1 + 2(X_r - \mu)^2/\sigma^2)\ \exp\left[\frac{3}{2}\ (X_r - \mu)^2/\sigma^2\right]$$

$$X_r^{1V} = (2\pi)^2(X_r - \mu)(7 + 6(X_r - \mu)^2/\sigma^2)\exp\left[2(X_r - \mu)^2/\sigma^2\right]$$

Using these formulae we may calculate the approximate moments of $R_{(rs)}$. Thus by way of illustration suppose that we consider two distributions that are N(0,1) and N(3,1) and that we take $n_1=n_2=20$, r=3, s=18, we find that

$$E(R_{3,18}) \simeq 0.140$$

$$\text{st error } (R_{3,18}) \simeq 0.093$$

In order to confirm these calculations and to get an impression of the shape of the sampling distribution a simulation was performed in which 500 pairs of samples of size 20 were generated from these distributions. The simulation produced,

$$\text{mean } (R_{3,18}) = 0.141$$

$$\text{st error } (R_{3,18}) = 0.096$$

The shape is shown in figure 10.3 demonstrating that the non-normality is not sufficient to make the use of the first two moments misleading.

The sampling distribution becomes less symmetrical if one compares two distributions with unequal variances. Figure 10.4 illustrates the results of a simulation of N(0,1) and N(3,4) with sample sizes of 20. This time the theory predicts that $R_{(3,18)}$ will have an average value of -0.061 and a standard error of 0.123. The simulation gave corresponding estimates of -0.065 and 0.123

The effect of sample size can be seen in figures 10.5 and 10.6. In these figures the same distributions have been used but with sample sizes of 40 and r=6, s=35. The normality is noticeably better even in the case of unequal variance.

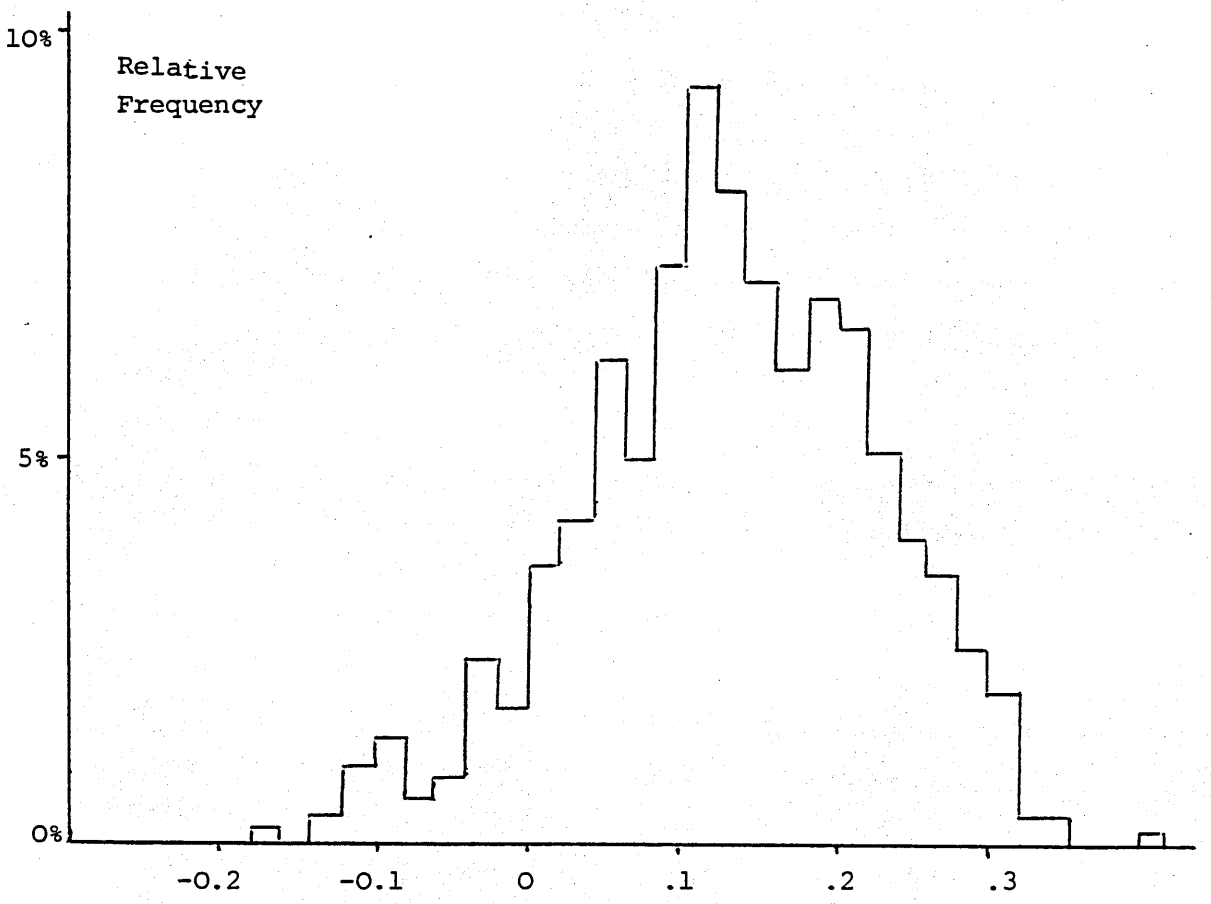


Figure 10.3

The Sampling distribution of $R_{3,18}$ for samples of size 20 from N(0,1) and N(3,1)

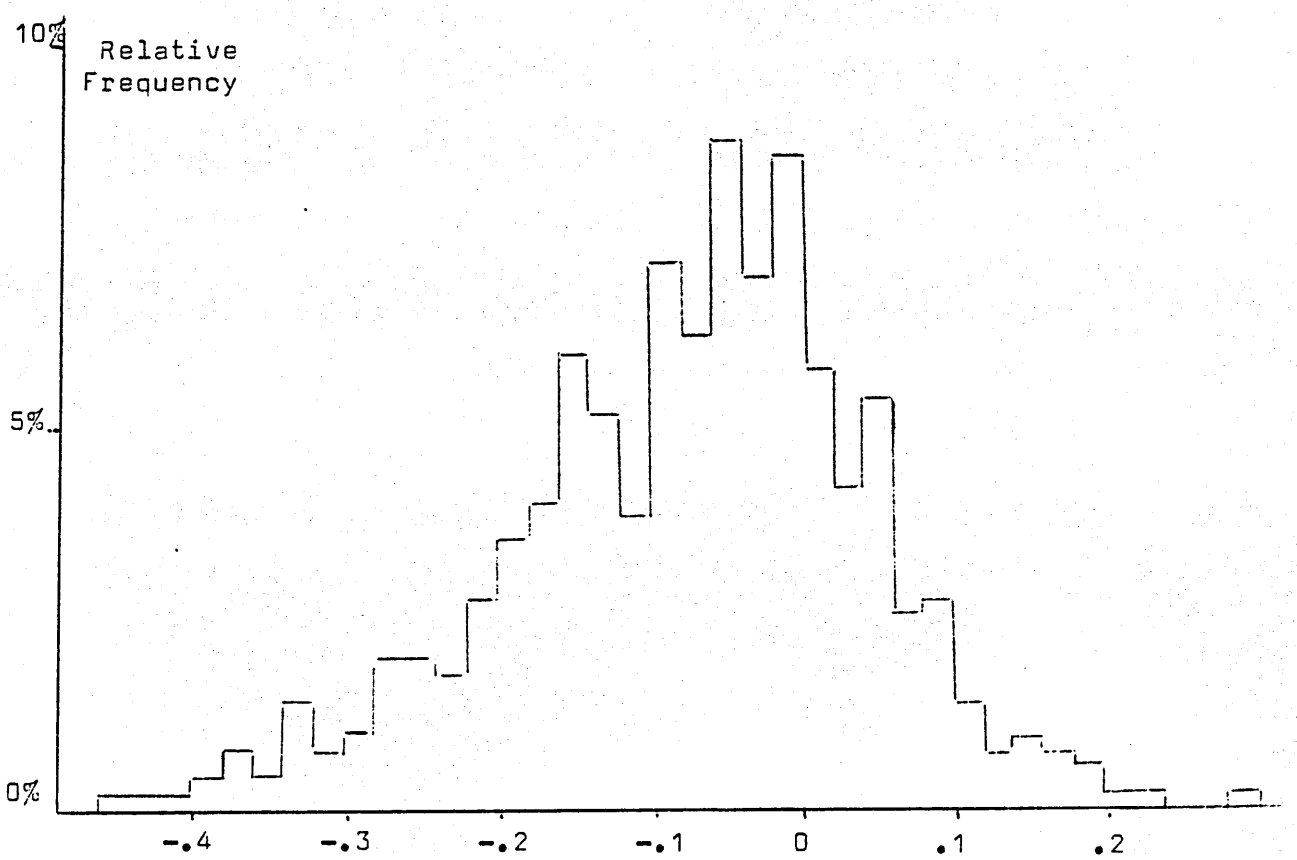


Figure 10.4

The sampling distribution of $R_{3,18}$ for samples of 20 from N(0,1) and N(3,4)

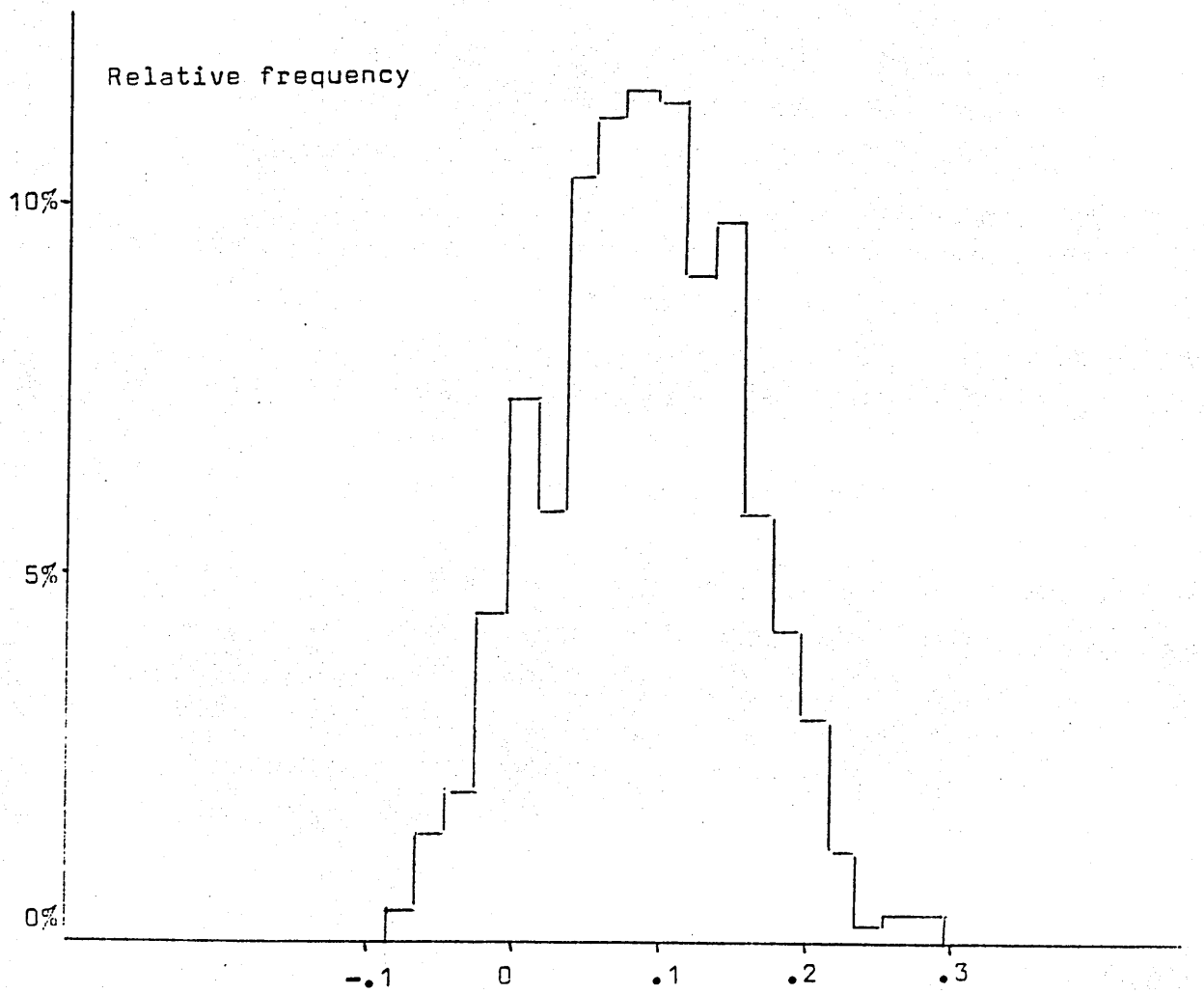


Figure 10.5

The sampling distribution of $R_{6,35}$ for samples of 40 from N(0,1) and N(3,1)

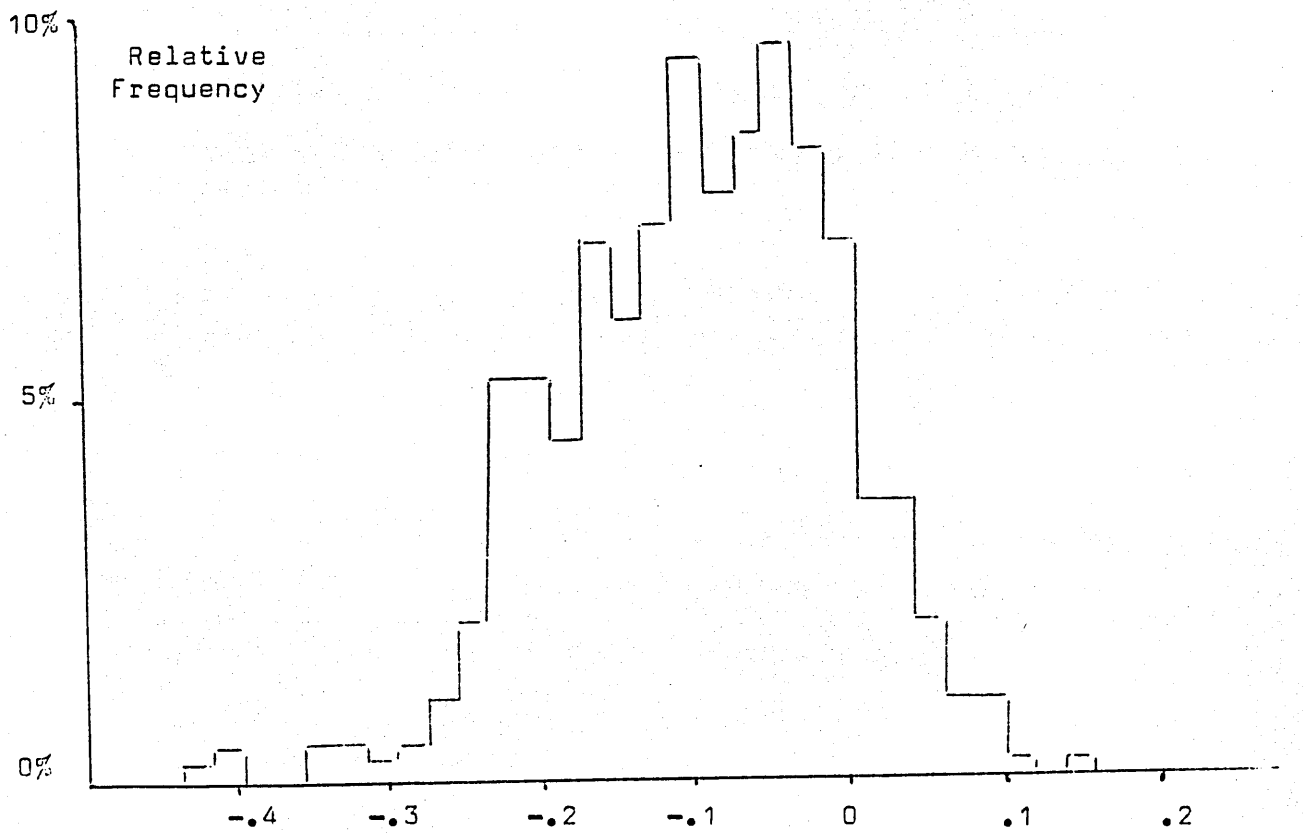


Figure 10.6

The sampling distribution of $R_{6,35}$ for samples of 40 from N(0,1) and N(3,4)

10.5 Comparing Measures of Separation

It is very difficult to obtain fair comparisons between measures of separation since their performance is bound to vary with the structure of the data. In the usual way we will look at the comparative performance of R and a commonly used alternative in the case where the data are normally distributed. The results do need to be interpreted with care as R is meant as a non-parametric measure and the alternative is not. These results are more illustrative of the type of analysis that could be performed given an actual problem.

The measure of separation that we will use in the comparison is one geared to normal data, that is,

$$D = \frac{\bar{x}_1 - \bar{x}_2}{\frac{1}{2}(s_1 + s_2)}$$

A simulation was performed measuring D for 500 samples of size 20 using the distributions employed in the previous section. The results were,

	D	sample mean	st.dev.
N(0,1) vs N(3,1)	3	3.070	0.480
N(0,1) vs N(3,4)	2	2.035	0.401

In themselves these figures tell us very little about the relative merits of R and D. However a comparison becomes possible if we look at their relative abilities to distinguish between two similar pairs of populations. Thus suppose that the parameters are changed slightly so that we were to compare N(0,1) with N(3.5,1), clearly D increases to 3.5, a change of 0.5 which is 104% of the standard deviation. The same change in R is from 0.141 to 0.215 only 77% of the original standard deviation. Clearly D is preferable for this configuration, which is as one might expect given that the data are normal.

Changing N(0,1) and N(3,4) into N(0,1) and N(3.5,4) produces an 83% change in D and a 63% change in R, again favouring D.

If the sample sizes are increased to 40 the results become;

	R	D
N(0,1) is N(3,1)	115%	150%
N(0,1) is N(3,4)	87%	113%

showing a similar advantage for D.

Once again the particular comparison is less important than the general method. In the next section we will use the idea of the sensivity of a measure of separation to changes in the underlying configuration in order to select the best values for r and s in the definition of $R_{(rs)}$.

10.6 The Sensitivity of R

The most important property of any measure of separation is that it can, on the basis of sampled data, tell reliably which of two configurations is the further apart. In this section we will consider how good R is in this respect. It is, of course, not possible to look at all of the situations that might arise in the use of a non-parametric measure, but the method is general and can be applied to any problem of special interest.

The simplest case, which we now consider, is when the classes are normally distributed, $N(0, \sigma^2)$ and $N(d, \sigma^2)$. Figure 10.7 shows, for n=20, the expected value of $R_{(3,18)}$ against the separation d/σ. It will be noted that $E(R_{(3,18)})$ varies much more quickly with small changes in the separation when the distributions are close together, this is as one would wish in practice, although the standard error of R is also an important consideration.

Because of the problems involved in sampling we cannot rely on the expected values alone but must also consider the accuracy with which the measure can be estimated. This may well also vary with the separation, as is shown in figure 10.8, which shows the standard error of $R_{(3,18)}$ plotted

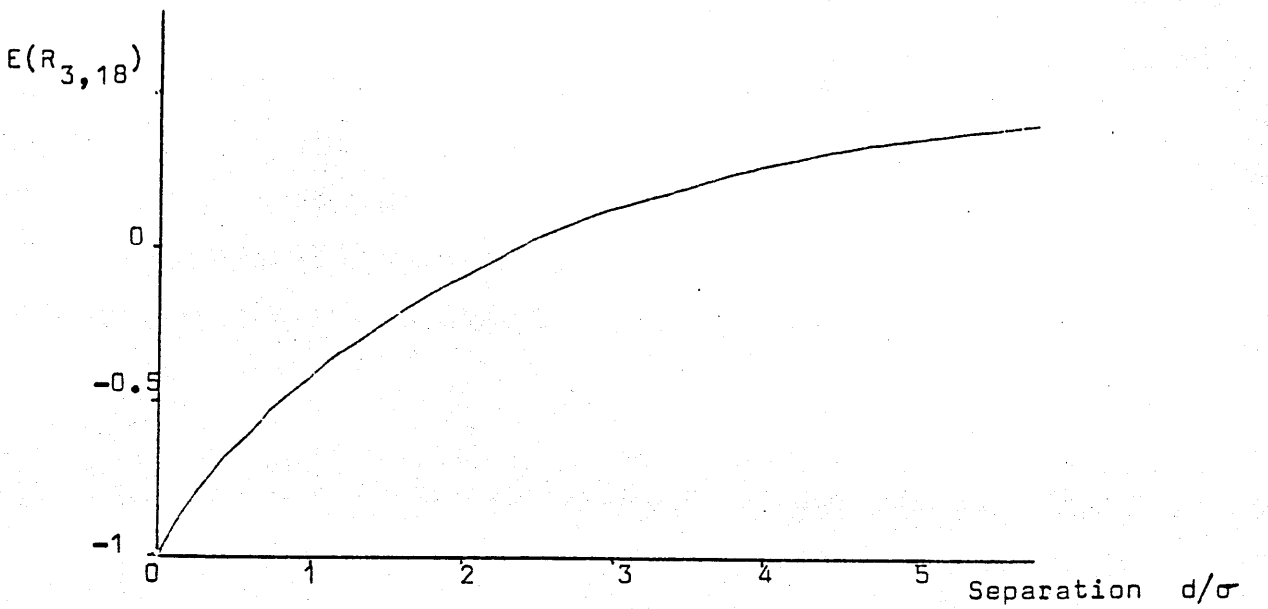


Figure 10.7
Variation in $E(R_{3,18})$ with separation

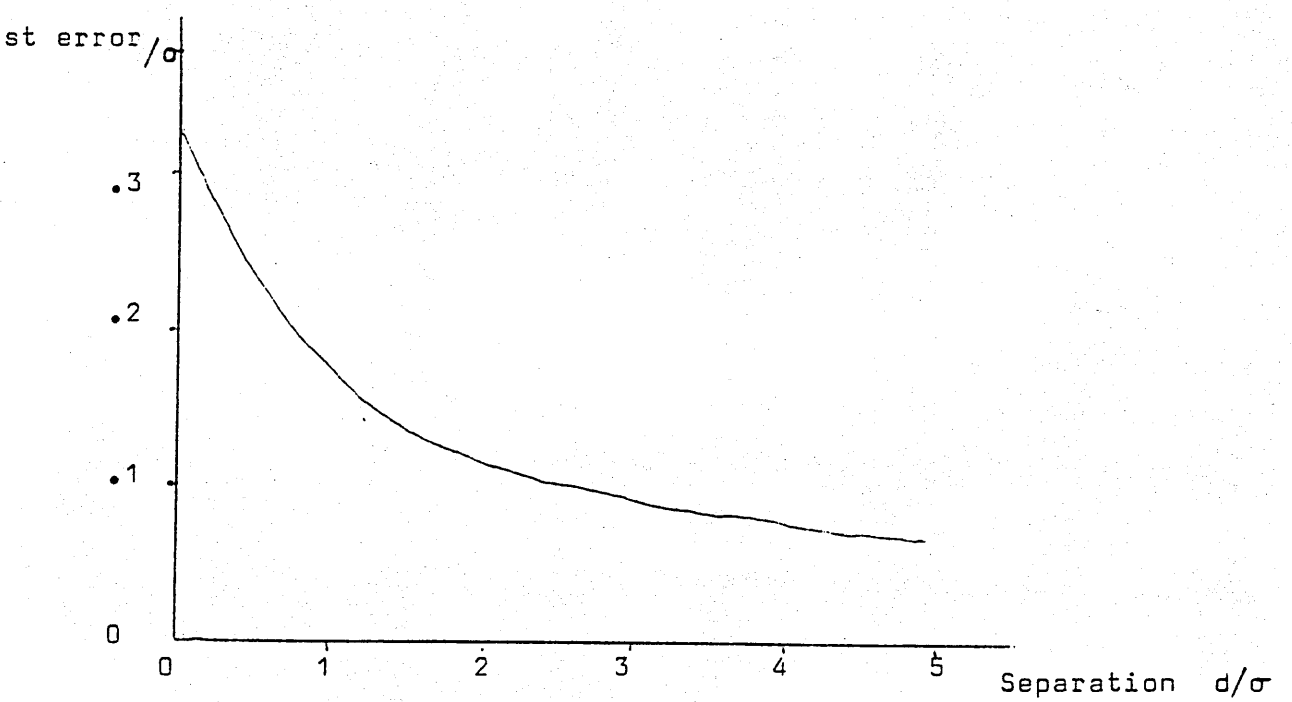


Figure 10.8
standard error of $R_{3,18}$ for various separations

against the separation.

To produce a sensitive measure of separation we require that a small change in the separation should produce a large change in the measure, that is large relative to the standard error with which the measure is estimated. We thus adopt the following measure of sensitivity,

$$R = \frac{dR}{dD} \Big/ \text{st error } (R)$$

We have already seen that this value will vary with the separation, however a graph of the type shown in figure 10.9, covering a practical range of separations gives a good overall picture.

In this particular case different pairs of order statistics have been compared. Thus we see that the use of the extremes from samples of size 20, i.e. the order statistics 1 and 20, performs particularly badly when the separation is small, say less than 2, and at wide separations the order statistics furthest from the extremes perform least well.

Were we required to choose the best order statistics on which to base out measure then clearly our answer would depend upon the separation, but either (2,19) or (3,18) would seem to be reasonable over most of the range.

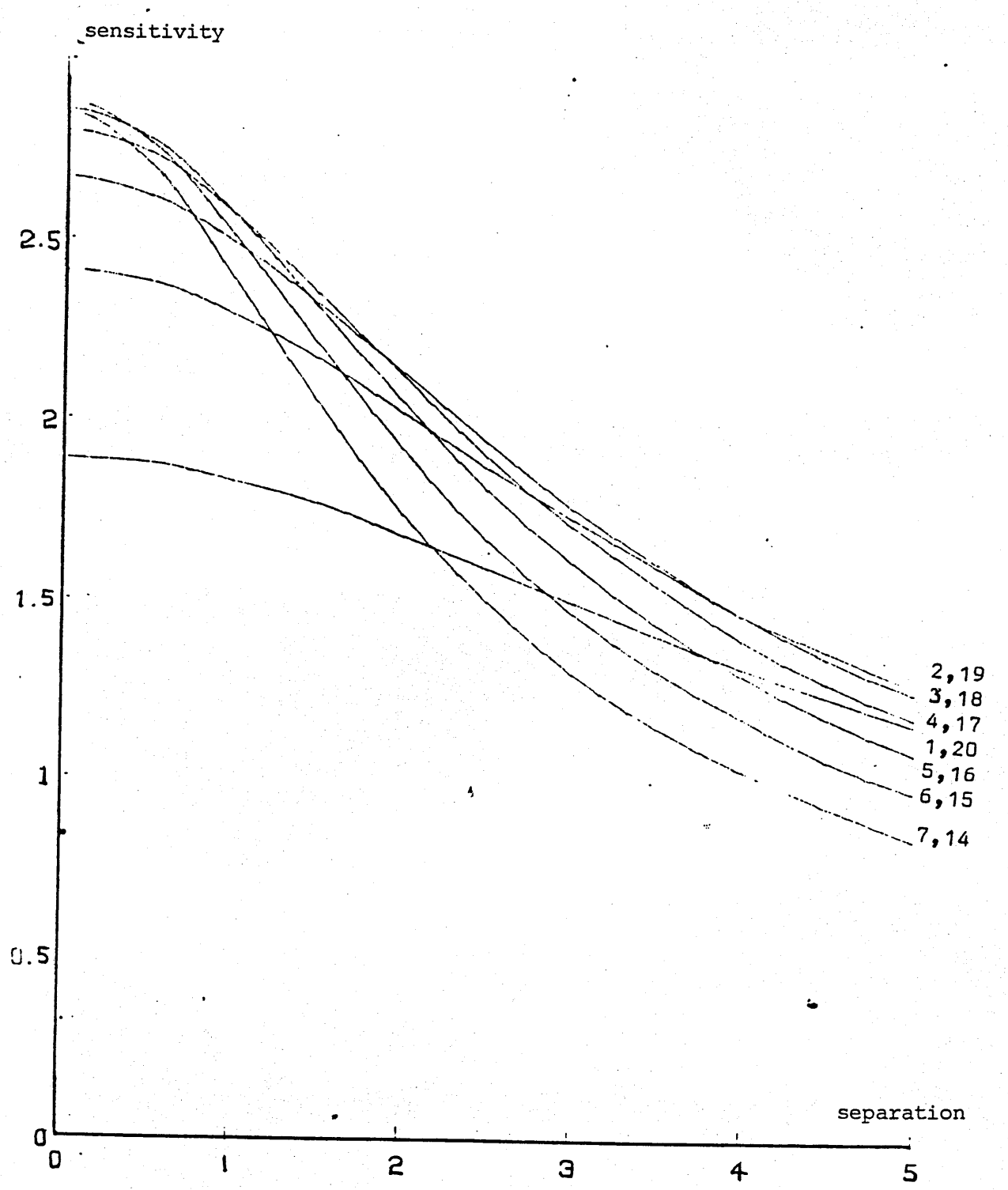


Figure 10.9

The sensitivity of R for various combinations of order statistics

10.7 Problems with Multidimensional Data

It must be born in mind when considering these results that they assume that one has a pair of univariate normal distributions. The normality assumption is not critical in that the methods used are quite general and could be repeated for other cases. However, the restriction to univariate distributions is bound to have an effect.

The results quoted so far would be applicable were R to be used as a univariate feature selection criterion but if the problem of selection or of classifier design is to be viewed in a multivariate way then the picture will change. In the next chapter R is used to define a linear classifier by searching for the linear combination of the observed variables that maximises the separation. Our results do not allow for the process of choosing an optimum and the sampling distribution found in section 10.4 cannot therefore be used. All that we can say is that,

(a) For large samples from multivariate normal populations with equal covariance structures, maximising R will lead to the usual linear discriminant function.

(b) The definition of R is non-parametric and should remain sensible so long as one guards against complete overlap.

11. Application of the Method of Non-parametric Discrimination

11.1 Introduction

We will now consider the application of the measure of separation suggested in the previous chapter to the analysis of the ballistocardiograms and the data on depression. Although the measure of separation was suggested so that it would remain sensible for any continuous distribution we have not as yet seen it applied to discrete problems.

The psychiatric data has symptoms measured on a four or five point scale and as such would appear to contradict the basic assumptions of a tolerance interval based approach. However, it is used to define linear combinations of the symptoms, so producing scales which although still discrete actually contain many values. The performance of the resulting classification scheme seems to suggest that under these conditions the method is still giving meaningful results.

11.2 The Ballistocardiograms

In this section we will look at the data on ballistocardiograms described in detail in chapter three. Following the feature selection described earlier it would appear that of the wave pattern the important variables are the amplitudes of the H,K,L,M and N waves; these are the features used in this analysis.

The algorithm used merely consisted of a numerical search for the linear function of the five variables that gave the largest measure of separation R. Experience showed that it was necessary to try several different starting points before one could be sure of converging to the maximum separation. The optimising algorithm used was that described by Nelder and Mead(1965). This algorithm is relatively quick and does not require that the derivatives of the function be known, indeed it does not even require that the function be free of discontinuities and as such is well suited to this type of problem.

The sample sizes vary in this data set, there being 50 pathological and 81 normal waves, consequently if one wishes to achieve the same approximate quantile for each distribution it is necessary that we use different order statistics. In the first analysis the value of p was kept

close to 0.05 by using the 3rd and 48th order statistics from the pathological data set and the 4th and 78th from the normal group. Using these values one finds that the optimum separation is R=-0.081, that is an overlap of 8% of the combined range. The linear combination that gives this separation being described by the weights,

H	K	L	M	N
.30	-.90	-.19	-.24	.12

The amplitudes have roughly comparable variability and so the weights give a good indication of the relative importance of each wave. It should be noted that this solution differs from that obtained in chapter nine.

Figure 11.1 shows the scores for each case on this scale, and shows immediately the highly fragmented nature of the distributions. As has been mentioned before this is almost certainly due to the fact that we have repeat measurements on a few subjects.

Forced discrimination could be based on a threshold of,

$$\tfrac{1}{2}|x_{1(3)} + x_{2(78)}| = 81$$

Even within the training sets used to define the scheme this results in a 24% error rate.

A partial discrimination scheme might be based upon thresholds at $x_{1(3)}$ and $x_{2(78)}$ giving the required error rate of 5% but leaving 38% of cases unclassified.

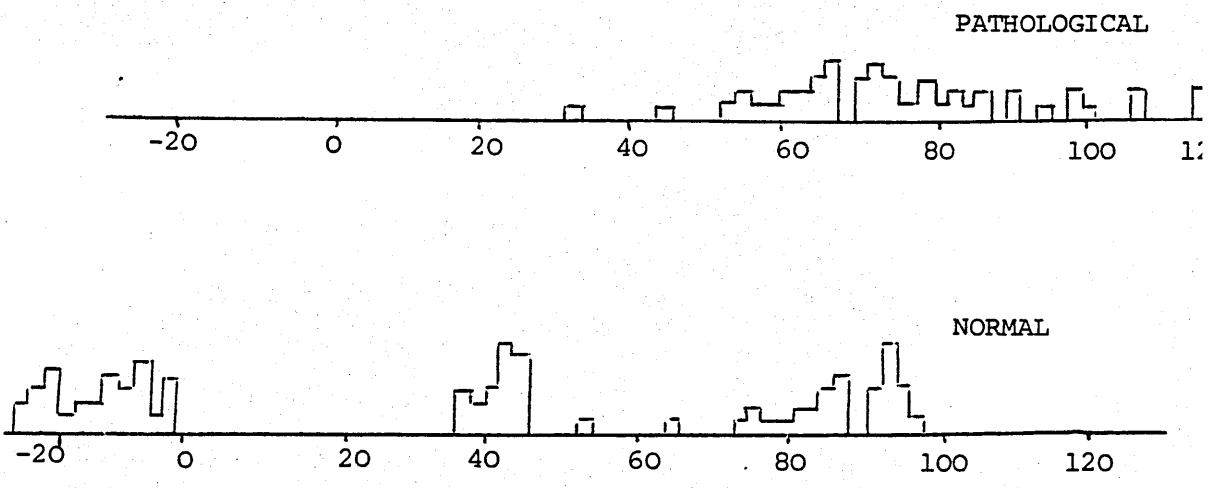


Figure 11.1

Individual scores for the linear combination of features that maximises R(0.05)

A second analysis was performed with the 1st and 50th order statistics from the pathological group and the 2nd and 80th from the normals. Thus achieving a p value of about 0.02. The solution changes quite dramatically now giving more weight to the H and M waves. The results are,

H	K	L	M	N
.40	-.20	-.25	-.84	-.15

The forced classifier gives an error rate of 24%, just as before, and the partial classifier gives the required 2% error rate and a 44% unclassified rate. The individual scores are shown in figure 11.2.

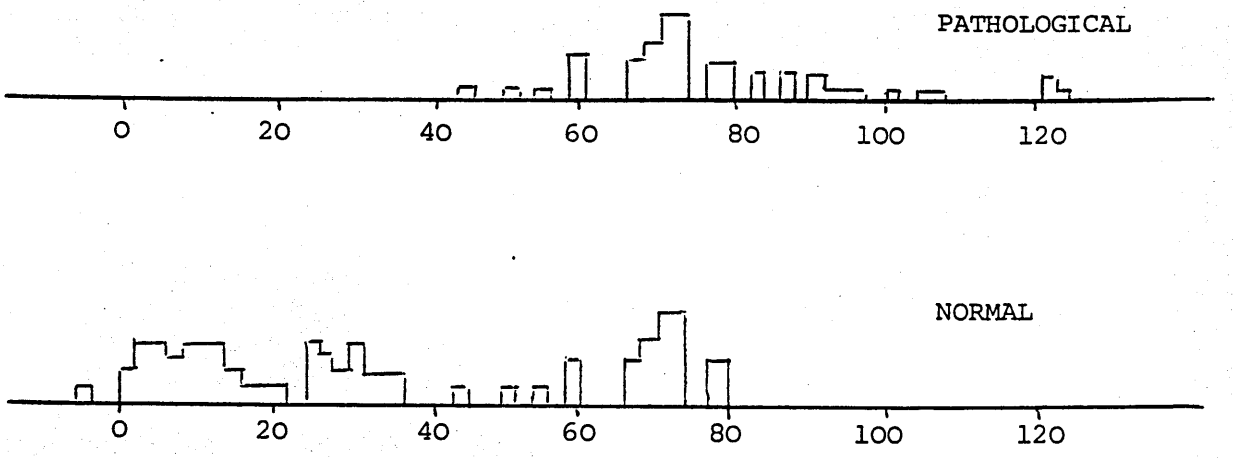


Figure 11.2
Individual scores for the linear combination of features that maximises R(0.02)

The difficulties encountered with this data set are due to it having been collected as repeated measurements on a few subjects. The fragmentary nature of the distributions making it very difficult to find a single linear classifier. There would be good grounds for treating these data as coming from a series of sub-populations, one for each individual, but the necessary information is not available.

11.3 The Psychiatric Data

This data set contains details of 42 symptoms on patients with one of five types of depression. In order to illustrate the use of the measure of separation, R, it was necessary to treat the classes in pairs and to perform some preliminary feature selection. Many of the symptoms carry no information at all that is relevant to a particular pair of classes and to reduce the computational load a first stage selection was employed whereby, for that pair of classes, the 22 symptoms with the smallest mean difference in scores were eliminated from consideration.

Working with the remaining twenty symptoms an interactive program was used that enabled one to remove symptoms and to see the effect that that removal would have on the optimum separation. The program was used to reduce the number of symptoms by a sequential backward selection process stopping just before R went negative.

The sizes of the training sets vary quite considerably, the largest being 41 and the smallest being 15, thus some allowance needs to be made in the choice of the measure of separation. In the analysis that follow order statistics were chosen to keep as close as possible to $p=0.075$ and $q=0.925$, where this was not possible exactly the program

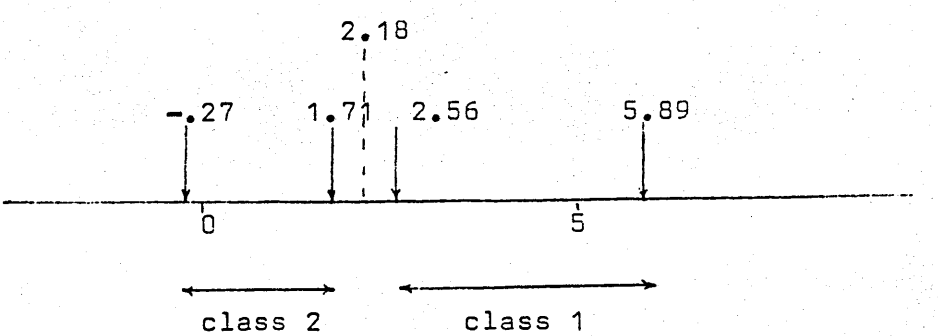
interpolated between the two surrounding order statistics.

SUMMARY OF RESULTS

CLASS 1 vs CLASS 2

Maximum separation R = 0.15

	Selected symptom	Weight
19	Hopelessness	.225
28	Loss of weight	.421
29	Delayed sleep	-.226
39	Delusion or hallucination	.849

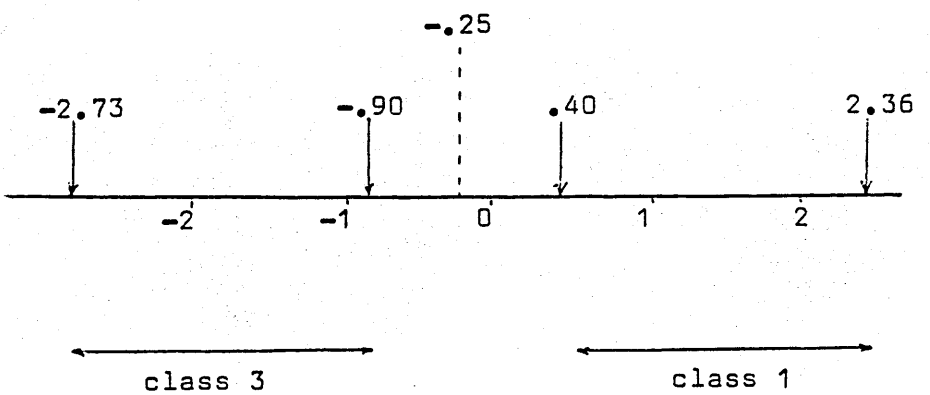


Class 1 is associated with high scores on symptoms 19, 28 and 39 but a lower score on symptom 29

CLASS 1 vs CLASS 3

Maximum separation R = 0.26

	Selected symptom	Weight
16	Brooding	-.326
18	Depressed mood	-.374
32	Irritability	-.323
38	Observed depression	.473
39	Delusion or Hallucination	.653



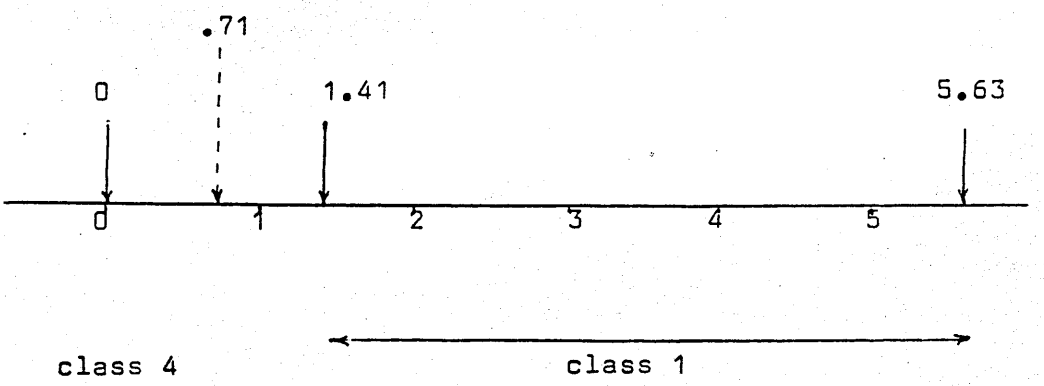
As with the comparison between classes 1 and 2, class 1 is again best distinguished by a high score on symptom 39.

CLASS 1 vs CLASS 4

This is a special case in that subjects in class 4 always score zero on symptoms 36 and 39, whereas subjects from class 1 always score more than zero. The fact that R is less than one reflects the variability within class 1.

Maximum separation R = 0.25

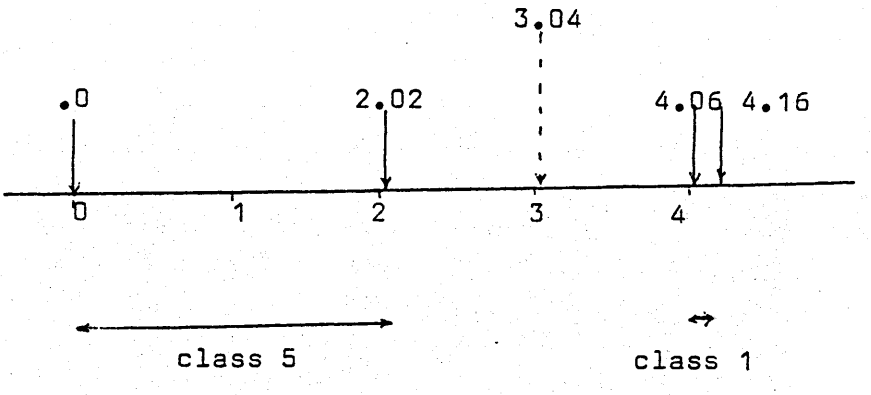
	Selected symptom	Weight
36	Agitation	.707
39	Delusion or hallucination	.707



CLASS 1 vs CLASS 5

Maximum separation R = .49

	Selected symptom	Weight
18	Depressed mood	.999
36	Agitation	.011
38	Observed depression	.019
39	Delusion or hallucination	.022

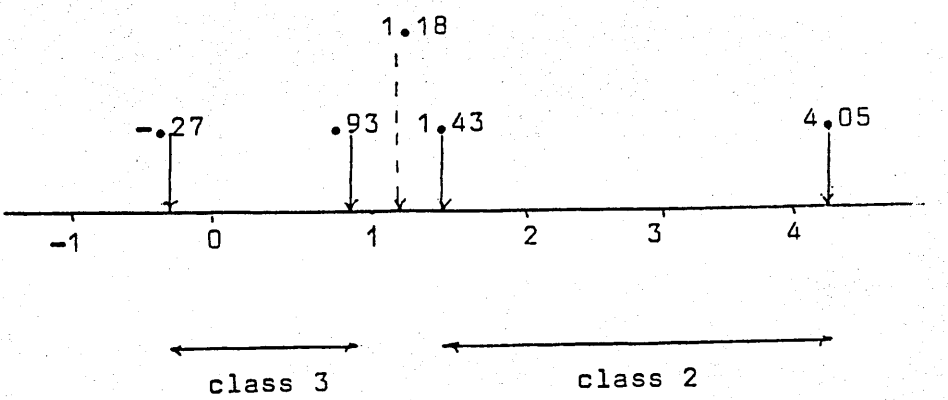


Clearly discrimination is based almost entirely on the observation of a depressed mood. As with the other comparisons involving class 1, delusion and hallucination are still good indicators, but of relatively less use in this case.

CLASS 2 vs CLASS 3

Maximum separation R = .11

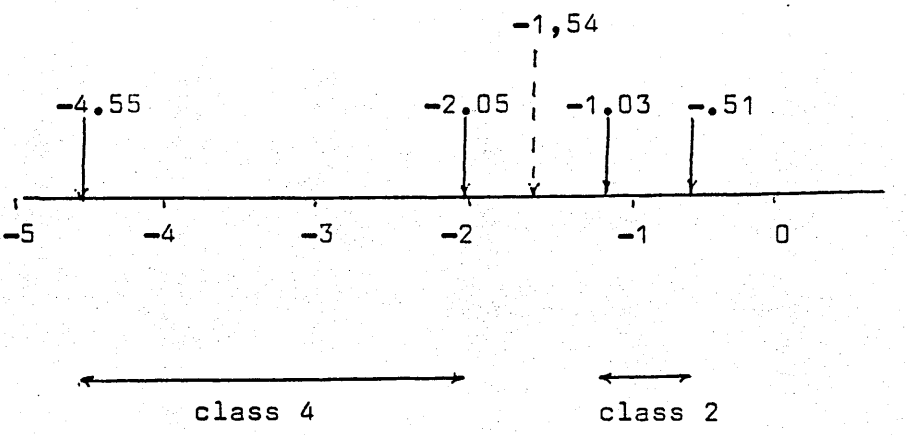
	Selected symptom	Weight
12	Specific phobias	.134
18	Anxiety avoidance	.312
25	Ideas of reference	-.156
28	Loss of weight	-.094
35	Subjective loss of affect	.167
36	Agitation	.828
38	Observed depression	-.119
41	Obsessive symptoms	.351



CLASS 2 vs CLASS 4

Maximum separation R = .25

	Selected symptom	Weight
18	Depressed mood	-.256
36	Agitation	-.819
41	Obsessive symptoms	-.513

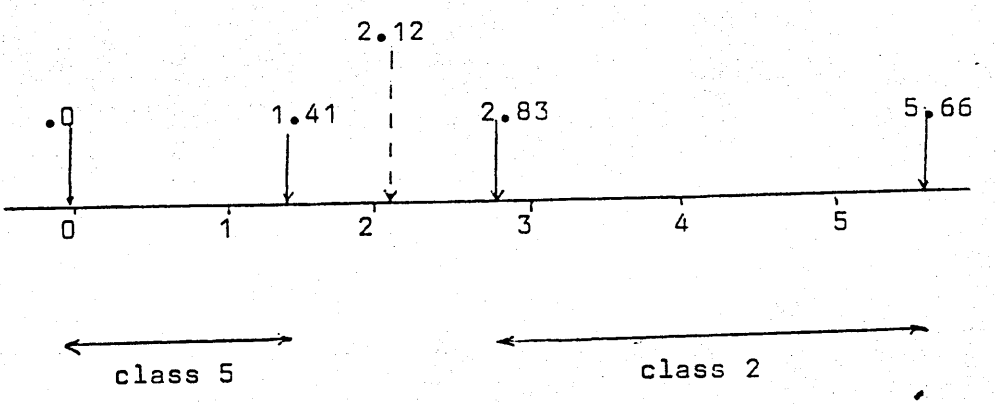


CLASS 2 vs CLASS 5

This is another special case in that the maximum score of any patient from class 5 on symptoms 18 and 36 is one, whereas the minimum score for a patient from class 2 is two.

Maximum separation R = .25

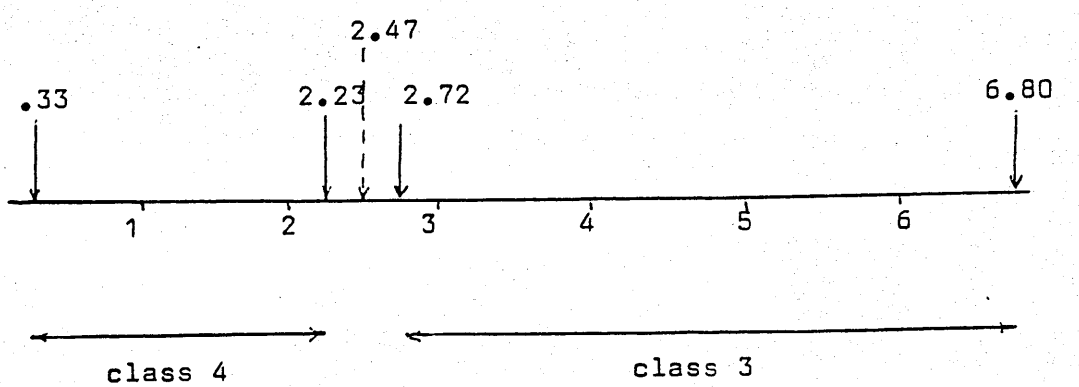
	Selected symptom	Weight
18	Depressed mood	.707
36	Agitation	.707



CLASS 3 vs CLASS 4

Maximum separation R = .08

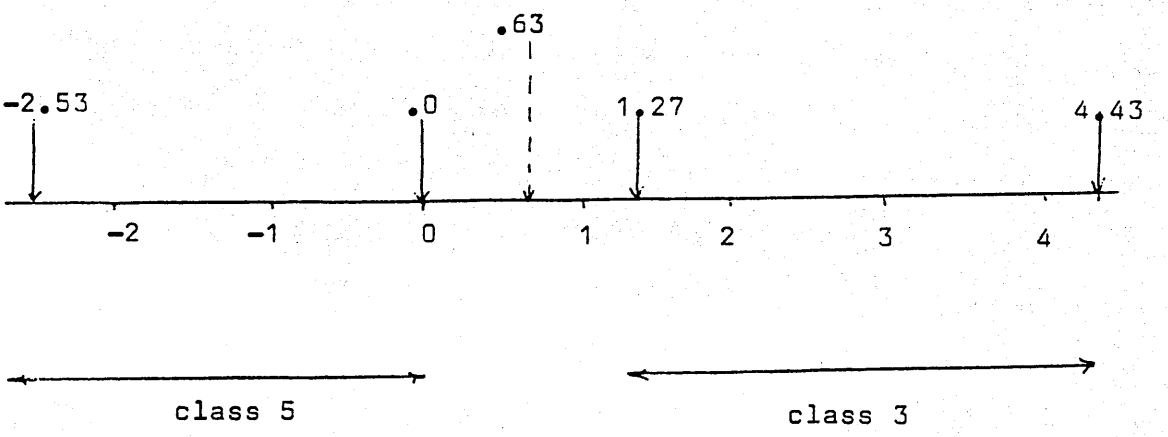
	Selected symptom	Weight
2	Tension pain	-.119
10	Situational autonomic anxiety	.274
11	Autonomic anxiety on meeting people	.399
15	Poor concentration	.120
21	Morning depression	.324
27	Pathological guilt	.705
31	Early waking	.135
32	Irritability	.187
35	Subjective loss of affect	.286



CLASS 3 vs CLASS 5

Maximum separation R = .18

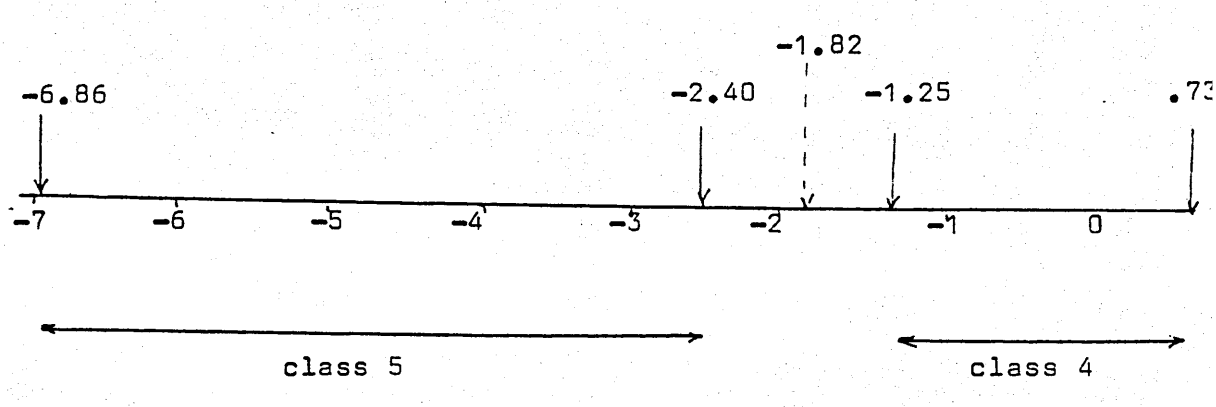
	Selected symptom	Weight
7	Nervous tension	-.316
21	Morning depression	.063
37	Observed anxiety	-.316
38	Observed depression	.633



CLASS 4 vs CLASS 5

Maximum separation R = .15

	Selected symptom	Weight
10	Situational autonomic anxiety	-.160
11	Autonomic anxiety on meeting people	-.609
15	Poor concentration	-.335
18	Depressed mood	.096
37	Observed anxiety	-.640
38	Observed depression	.270



In order to judge the worth of the scheme one must obtain some estimate of the error rate. In the absence of a parametric model for the data the best way of obtaining such an estimate is by using the 'leave-one-out' method suggested by Lachenbruch and Mickey(1968). According to this method each case is omitted in turn and the discriminant function is calculated using the remaining cases. The case that was left out is then classified according to this discriminant function and the results are averaged over all cases.

Since we have five classes of depression it is necessary to calculate all ten possible functions and then classify the omitted case according to each. If the subject is placed in one class in preference to all others then that is taken as its classification. If, on the other hand, there is a contradiction then the case is left unclassified; thus we have a type of partial discrimination scheme. By way of illustration table 11.1 shows the classification of two cases from class 2. In the first example class 2 is consistently preferred, but in the second there is no overall preference and so the case is left unclassified.

Table 11.1

Two examples of the classification procedure

	Classification	
Comparison	Example 1	Example 2
1 vs 2	2	2
1 vs 3	3	3
1 vs 4	1	1
1 vs 5	5	1
2 vs 3	2	2
2 vs 4	2	4
2 vs 5	2	2
3 vs 4	3	3
3 vs 5	5	3
4 vs 5	5	4
Result	Classify as 2	Unclassified

The results obtained when this classification was applied in a leave-one-out experiment are summarised in table 11.2. It will be noted that 126 of the 146 cases that the psychiatrist felt able to classify were correctly classified by this scheme, an overall success rate of 86.3%; of the remainder a further 10.3% were wrongly classified and 3.4% were left unclassified.

Table 11.2

The results of a 'leave-one-out' analysis

Psychiatrist's Classification	Classification Obtained 1	2	3	4	5	Unsure	Total
1	14	1	0	0	0	0	15
2	0	28	1	0	0	2	31
3	0	1	16	2	1	0	20
4	0	0	4	30	2	3	39
5	0	0	3	0	38	0	41
Unsure	0	0	1	1	2	0	4
Total	14	30	25	33	45	5	150

Expressing the results from table 11.2 as a percentage of the total numbers in each class we can see immediately, as in table 11.3, that the bulk of the misclassification is due to the uncertainty over classes 3 and 4.

Table 11.3

Percentage error rates from the 'leave-one-out' analysis

Psychiatrist's Classification	Classification Obtained					
	1	2	3	4	5	Unsure
1	93	7	0	0	0	0
2	0	90	3	0	0	7
3	0	5	80	10	5	0
4	0	0	10	77	5	8
5	0	0	7	0	93	0

12. EXTENDING THE METHOD OF NON-PARAMETRIC PARTIAL DISCRIMINATION TO HIGHER DIMENSIONS

12.1 Introduction

We saw in chapters ten and eleven how it is possible to define a univariate measure of separation that works well as the basis of a partial discrimination scheme. It might well be the case however, that in order to discriminate between two classes we need more than one dimension. In such a case it would be necessary to extend the definition of the measure of separation, and it is this problem that we consider in the present chapter.

The original measure uses the order statistics of the samples. Unfortunately there is no one obvious way of extending the idea of an order statistic into more than one dimension and it is necessary to work with a convex hull of extreme points. Calculation of the convex hull in many dimensions can be computationally difficult and currently available methods restrict their attention to two or three dimensions. However a new algorithm is described that makes the computation of the complex hull in any number of dimensions a practical proposition.

12.2 Multivariate Ordering

Bartlett(1976) in a review of multivariate ordering distinguished between four possible extensions of the univariate concept.

1. Marginal ordering

   When the data are ordered on the value of one marginal distribution or perhaps, on a univariate variable formed by combining multivariate measures.

2. Reduced ordering

   When the ordering is based on some multivariate measure of distance, such as that of Mahalanobis.

3. Partial Ordering

   When the data are divided into subsets. One set consisting of 'extreme points', another of 'less extreme points' and so on.

4. Conditional Ordering

   When one successively orders the data according to different marginal distributions. Thus one might form a set of extreme points according to one variable, and then order them according to the values of a second variable.

Although the divisions between the four methods are not always clear cut they still form a useful breakdown of the possible approaches to the problem.

The method that is best suited to the extension of

partial discrimination into higher dimensions is the partial ordering resulting from the convex hulls of the sample. The outermost convex hull, in two dimensions, is defined as the convex covering polygon with minimum area; this is intuitively equivalent to imaging the points in the sample as nails in a board, the outermost convex hull would then be obtained by wrapping an elastic band around the nails. Successive convex hulls may be obtained by removing the points in the outermost hull, then repeating the process with the remainder; extensions to higher dimensions are obvious. Figure 12.1 shows the convex hulls of one sample.

The convex hulls of a sample have been used occassionally in statistics especially in producing estimates of the correlation between two variables. For a review of the applications of convex hulls to pattern recognition see Toussaint(1978).

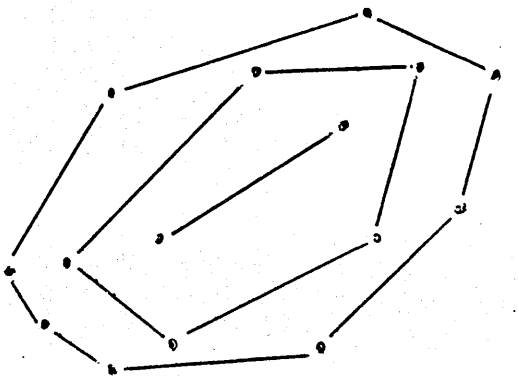


Figure 12.1

An example of the convex hulls of a sample

12.3 Bivariate Partial Discrimination

Kendall(1966) suggested the simple idea that two convex hulls A_1 and A_2, as illustrated in figure 12.2, could be used for non-parametric classification by using the rule,

if $x \in A_1 \cap A_2'$ classify as w_1

if $x \in A_1' \cap A_2$ classify as w_2

else leave unclassified

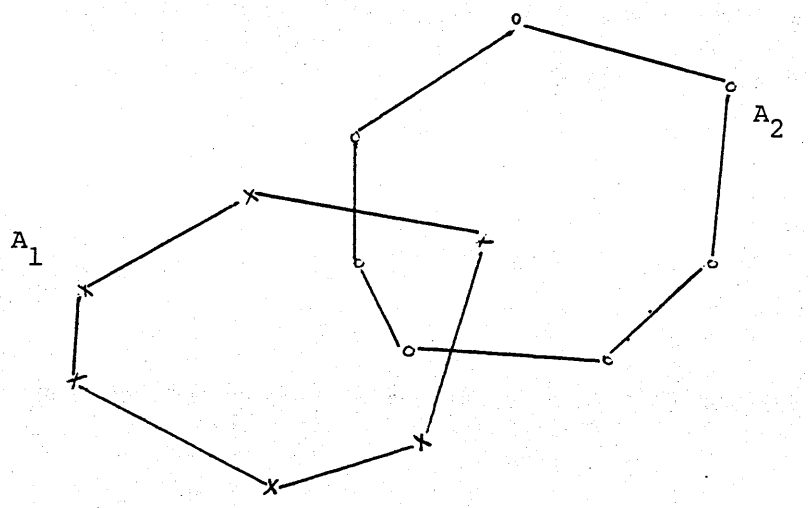


Figure 12.2

Partial classification using two convex hulls

The problem that remains is to find the plane within which the best pair of convex hulls lies when one has more than two measurements. The measure of separation used in chapters ten and eleven employs the overlap between the two distributions. A natural extension would be to use,

$$R = \frac{\text{Area } (A_1 \cap A_2)}{\text{Area } (A_1 \cup A_2)}$$

Thus we would attempt to find the two orthogonal linear combinations of the available variables which define the plane in which the ratio R is as small as possible. Once again taking care to avoid complete overlap.

Unlike the univariate measure this measure does not extend to completely separated populations.

Because of the heavy computational load involved in calculating the convex hull it may be preferable, especially if there are many variables, to find the first linear combination by the methods of chapter eleven. That combination could then be held fixed whilst one searches for the best orthogonal direction, the pair defining the best plane.

Extensions to three and higher dimension are obvious with the use of,

$$\frac{\text{Vol } (A_1 \cap A_2)}{\text{Vol } (A_1 \cup A_2)}$$

12.4 Calculating the Convex Hull

During the 1970's a number of algorithms were proposed for finding the convex hull of a set of points. However most restricted their attention to points in a plane. Graham(1972) suggested that points in a plane might be expressed in polar co-ordinates relative to some interior point. The points can then be sorted according to their angles, and searched through in threes. Graham gave a method for eliminating those points not on the hull.

Jarvis(1973) advocated an approach similar to that which would be used if the hull were to be drawn by hand. Thus the leftmost point v_1 is found and the angles of all other points relative to a vertical through v_1 are found. The largest angle zv_1v_i gives the next vertex v_i. This is then used as the starting point for the next stage. See figure 12.3.

Recursive algorithms have been suggested by Preparta and Hong(1977) and by Bentley and Shamos(1978). Preparta and Hong show how to merge two convex hulls so as to form the convex hull of the combined set, after the fashion of figure 12.4.

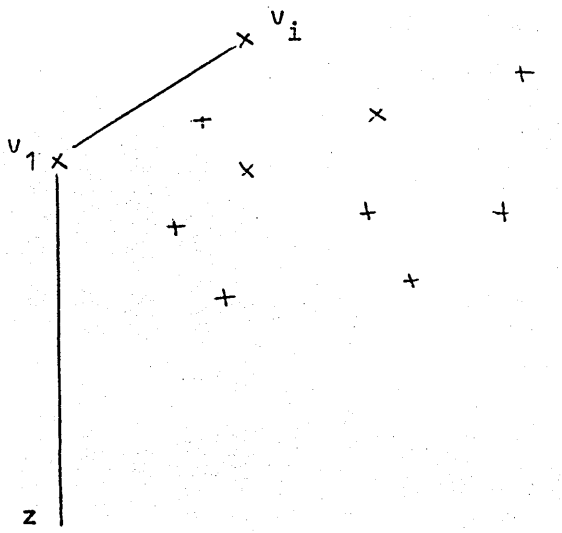


Figure 12.3

Initial calculation in Jarvis's method

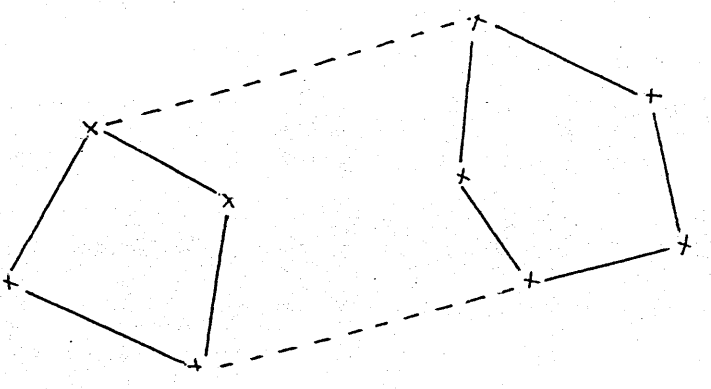


Figure 12.4

Merging two hulls in the method of Preparta & Hong

The resulting recursive algorithm is then;

Input: a set S of points $(a_1, a_2, \ldots a_n)$ sorted by one co-ordinate.

Step 1: divide S into sets of 1,2 or 3 points so that each set forms its own convex hull.

Step 2: merge the sets in pairs, forming the merged convex hull, continuing until a single hull is formed.

They also show how their algorithm can be extended to three dimensions, giving details of a method for merging two three dimensional hulls.

Green and Silverman(1979) described an algorithm, again restricted to the plane, which itself is an improvement on an earlier algorithm of Eddy(1977). Their method requires that two extreme points v_1 and v_2 be found; these must by definition lie on the final hull. If all points lie on one side of v_1v_2 then it is an edge of the final hull. If not, locate the most extreme point in the direction perpendicular to $v_1 v_2$, say v_3 and v_4. Lines v_1v_3, v_1v_4, v_2v_3 and v_2v_4 are then tested in the same way. The algorithm stops when all potential edges have been tested.

Green and Silverman consider some of the numerical problems that arise in the application of their algorithm. The main difficulty being with nearly collinear points, together with the finite word length of a computer. This problem is considered again in the next section

One algorithm that does consider the general problem of the convex hull in p-dimensions is the little quoted paper of Chaud and Kapar(1970). Their idea was that since each edge of a p-dimensional hull is the intersection of two faces, it must be possible to start with an edge and a face, to rotate the face about the edge to form a second face. The extreme point(s) in that plane will define a new set of edges and the process may be continued until the whole hull is found.

Comparing these various algorithms Green and Silverman found theirs to be the quickest at finding a hull within a plane.

This general problem has been shown by Avis(1979), to have a complexity of O(nlogn), where n is the number of points. The pattern recognition literature has however been much concerned with a related problem, that of finding the convex hull of a polygon. Since a polygon has more structure than a general set of points it is not surprising that a simpler solution is possible and Sklansky(1972) suggested the first O(n) algorithm for the problem. This algorithm has basic similarities with Graham's method and has attracted a

lot of attention. Bykat(1978) showed that under certain circumstances the algorithm would fail and Orlawski(1983) and Goshi and Shyamasunder(1983) amongst others, have described modifications. Unfortunately there is no generalisation of this method out of the plane.

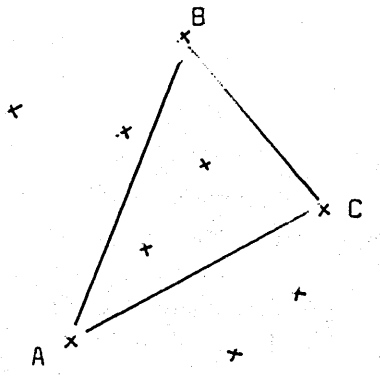
12.5 A New Algorithm for Locating the Convex Hull of a set of points in p-dimensions

This algorithm requires a set of p+1 distinct points which together form the starting point for the calculation of the hull. The p+1 points must be chosen so that they are extreme in distinct directions thus they all lie on the final hull and it is advantagous to the speed of computation if they are widely spread.

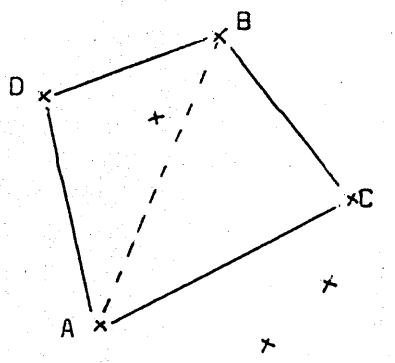
The set of p+1 points will form a p-dimensional tetrahedron. Points lying inside this tertrahedron may be ignored in all future calculation since they cannot possibly lie on the final hull.

Each face of the tetrahedron is considered in turn. If no points lie outside the face then it must be a face of the final hull. If, however, there are points outside the face then the most extreme, measured in a direction normal to the face, is located.

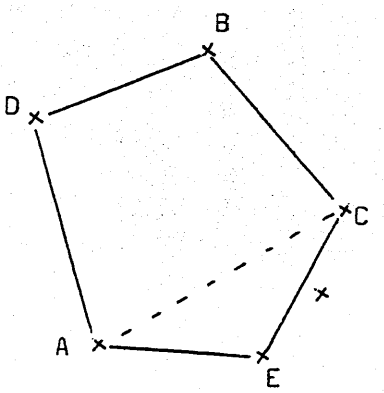
This extreme point must itself lie on the hull. Together with the p points from the face under test, this new extreme point forms a p-dimensional tetrahedron, p faces of which might be faces of the final hull. Once more any points inside the new tetrahedron may be ignored in future calculation as they cannot possibly lie on the final hull.



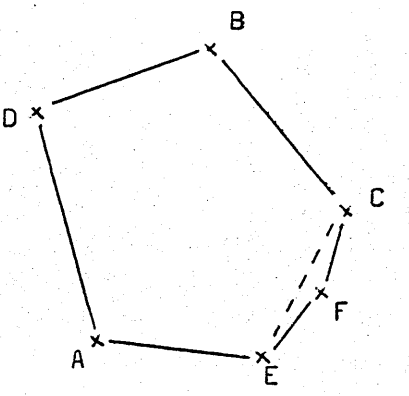
stage (i)



stage (ii)



stage (iii)



stage (iv)

Figure 12.5

An example of the use of the new algorithm for finding a convex hull in two dimensions

In this way new potential faces are generated and the algorithm continues until either there are no untested faces or until there are no remaining points, that is to say points that are neither known to lie on the hull nor eliminated.

The two-dimensional version of this algorithm is illustrated in figure 12.5.

The initial set of three points is ABC, together these three form a triangle. The points inside ABC may be ignored.

The further point outside AB is D. ADB is another triangle and AD and DB are potential faces for the final hull. Points inside ADB may be ignored.

E is the most extreme point outside AC and F is the most extreme outside EC. At this stage all points have either been eliminated or incorporated into the hull and consequently the computation is complete.

It will be seen that in two dimensions the calculations required are essentially the same as those needed for Green and Silverman's algorithm, except that this method progressively removes points that lie inside the hull. For this reason the algorithm will be no slower than Green and Silverman's and may be considerably quicker if a good initial tetrahedron is found and if there are a large number of points.

The major calculation in the algorithm is the score of a point normal to a particular face. If the face contains

points $(x_{11},\ldots,x_{1p})$, $(x_{21},\ldots,x_{2p})$ to $(x_{p1},\ldots,x_{pp})$. Then the normal is given by the determinant,

$$D = \begin{vmatrix} x_1 & x_2 & \ldots\ldots & x_p & 1 \\ x_{11} & x_{12} & \ldots\ldots & x_{1p} & 1 \\ \cdot & & & & \\ \cdot & & & & \\ x_{p1} & x_{p2} & \ldots\ldots & x_{pp} & 1 \end{vmatrix}$$

By placing the co-ordinates of the point to be tested into positions $x_1,x_2,\ldots x_p$, we obtain a score which will be positive on one side of the face and negative on the other. Further D increases in magnitude as the distance from the face increases.

In order to check whether or not the point is on the interior or exterior side of the face, one may compare its sign with that of the other point of the tetrahedron.

The one remaining problem is to find the initial set of (p+1) points. The simplest way to do this is to choose an initial direction and to take the two extremes in that direction. Consider then the distance of each remaining point, normal to the line joining the first two points. The largest in magnitude gives the next point for inclusion in the initial set. Progressively a point is added why finding the largest distance normal to those already selected until p+1 have been obtained.

As a by-product this procedure makes the very necessary

test that the points do in fact lie in p-dimensions and not in some subspace. For if the points lie in less than p-dimensions we will reach a point where all of the distances, normal to the selected points, are zero. If this happens then one can work with the tetrahedron formed from those points already selected and the algorithm will work perfectly well all be it in the subspace.

As Green and Silverman note it is necessary to take steps to avoid the degeneracy caused by nearly collinear points and the finite precision of the computer's arithmetic. Thus, in two dimensions, using figure 12.6 for illustration, the danger is that because of a lack of precision one might locate v_k as the most extreme point outside $v_i v_j$ but then find that v_k lies outside $v_l v_j$.

To avoid this problem one may require that to lie off the plane being tested a score must be at least c, where c is some suitably chosen small value. Green and Silverman found by experiment that $c=10^{-5}$ was a reasonable value and this value has been used in our algorithm.

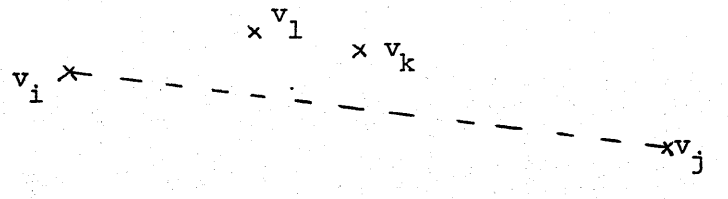


Figure 12.6

Illustrating the difficulties that can arise when calculating the convex hull due to near collinearity

12.6 Calculating the Volumes

The volume of the tetrahedron formed by p+1 points in p-dimensional space is given by D/p!, where D is the determinant calculated in the previous section. It is thus a simple matter to build up the volume of the convex hull by adding the volume of each tetrahedron as it is found.

The problem of the overlap is more complex. The difficulty is easiest illustarted in two dimensions. It will be seen from figure 12.7 that in order to find the overlap, one needs the convex hull of those points,

(i) from hull A_1 that lie inside hull A_2

(ii) from hull A_2 that lie inside hull A_1

(iii) points that lie on the intersection of the two hulls.

Whilst the location of the intersection points is only a matter of solving sets of linear equations the programming is quite complicated. It might well be easier to generate points at random from within one hull and find the proportion that lie within the other.

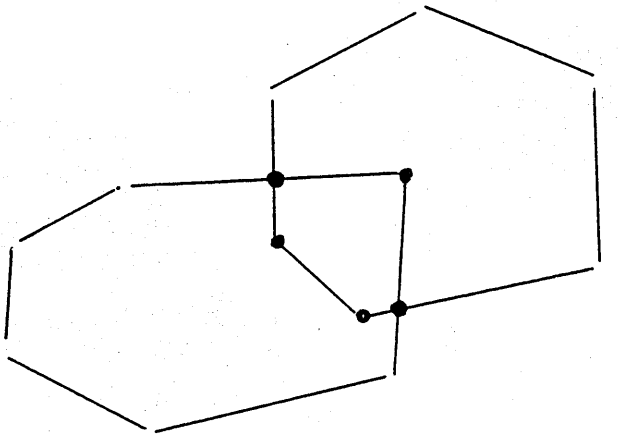


Figure 12.7

The points necessary to define the overlap of two hulls

The procedure for calculating the intersection points consists of finding all the faces of A_1 that have at least one point inside A_2 and at least one point outside A_2. Similarly one finds all faces of A_2 that lie partly inside and partly outside A_1. Each of the selected faces from A_1 then needs to be matched with each selected face from A_2. The matched faces will intersect along an edge. If that edge lies entirely within both hulls then it is an edge of the overlap.

12.7 Application to the Psychiatric Data

We saw in chapter eleven how it is possible to separate each of the five classes if they are treated in pairs. The problem that we consider here is slightly different in that we seek to discriminate between class five, the anxiety states, and the rest. Thus in the rest of this section classes one to four are treated as a single class.

The first stage in the analysis is to decide which of the features best distinguish between our two classes. This was done using the methods of chapter seven. The Euclidean distances between the five class means were used to define four princial co-ordinates and each case was assigned to the resulting four dimensional space on the basis of its distances from the five class means.

Having achieved this four dimensional representation we now need to place a plane through that space so that the two classes, five and the rest, are best separated. The first direction was found by the methods of chapter eleven, using the extremes of the classes to define the measure of separation R. The resulting weights were,

.832 -.081 -.009 .543

which give R=-0.140, an overlap of 14% of the combined range, as illustrated in figure 12.8

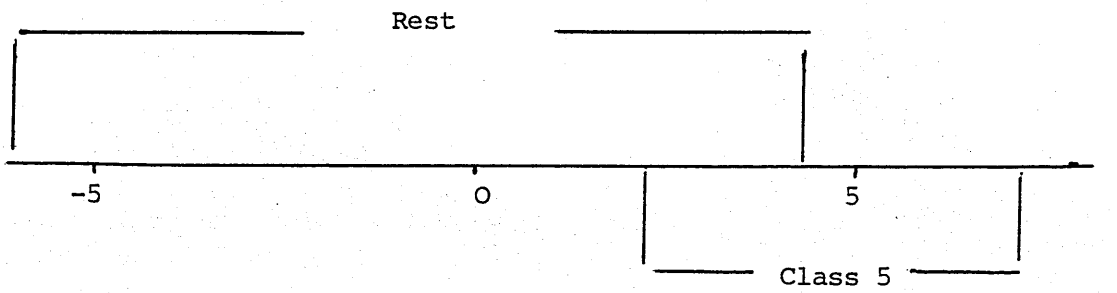


Figure 12.8

The best single feature for distinguish class 5 using the psychiatric data

Keeping this direction fixed the best orthogonal direction was sought, using the criterion of maximising,

$$R = \frac{\text{Area } (A_1 \cap A_2)}{\text{Area } (A_1 \cup A_2)}$$

In fact the best R that could be achieved was found to be -.057, an overlap of just under 6%. This was obtained when the second feature was defined by weights,

-.354 .267 .670 .596

The resulting solution is shown in figure 12.8. Given the pattern of symptoms of any future case it would be a relatively simple matter to place them in the four dimensional space and then locate them in this plane. The hulls could thus be used as the basis of a partial discrimination scheme.

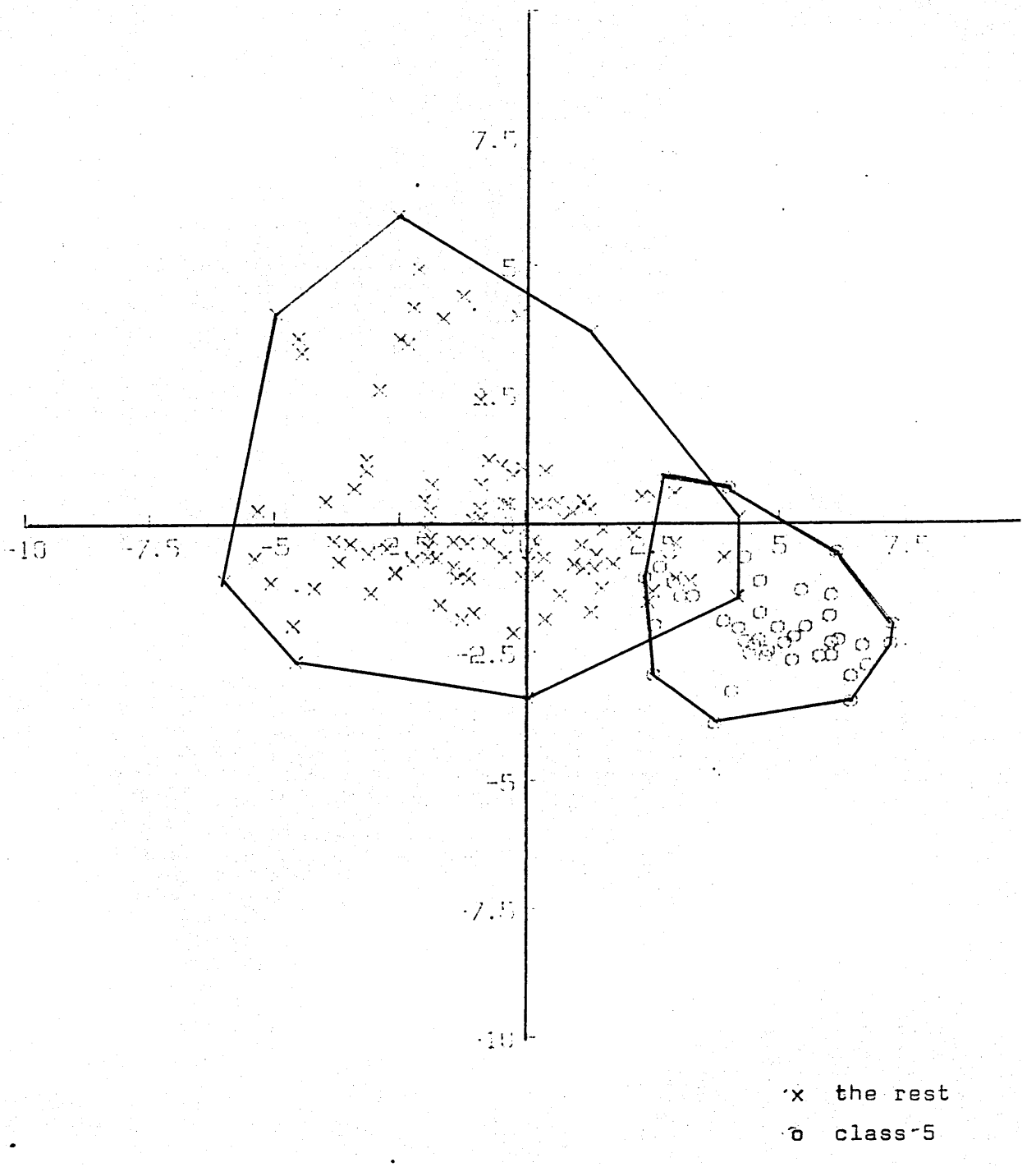


Figure 12.9

The two dimensional solution of the psychiatric problem

13. PATTERN RECOGNITION WITH IMPERFECTLY CLASSIFIED TRAINING SETS

13.1 Introduction

In section 13.2 we will see how one of the data sets that was intended for use in testing the methods of feature selection and classification that have been described in previous chapters, was in fact itself very poorly classified. This imperfect initial classification violates the assumptions implicit in almost all pattern recognition schemes and led to an investigation of methods that might be used to correct such errors. Whilst the initial misclassification in our data set proved so extensive that little could be done to correct it, these methods may prove useful for analysing less extensively misclassified data.

The problem of poorly classified training sets is probably more widespread than is generally recognised. Sometimes, as in medical diagnosis, it may not be possible to ensure correct initial classification, whilst in other fields of application perfect classification may be theorectically possible but prohibitively expensive.

When correct initial classification cannot be guaranteed it is important that we consider how any initial misclassification would affect the subsequent analysis and if possible we should use methods that are robust to such misclassification. Most of the published work on this

problem has, as we shall see in section 13.3, concentrated on the effect on the classifier, but it will, of course, also influence the feature selection.

One of the major problems in developing a realistic model for initial misclassification is that it is likely, in practice, to be those cases that lie near to the class boundaries that will be misclassified. Although this point is well appreciated almost all of the results relevant to this problem assume a random distribution of misclassification and as such must be viewed with care.

The measure of separation R, described in chapter ten, may be made robust by the choice of less extreme order statistics and it would certainly be advisable, if initial misclassification were suspected, to analyse the data using measures R based on a range of order statistics.

In this chapter we consider the ways in which two other standard methods of analysis can be made robust to initial misclassification by introducing a methods that could be applied to many schemes.

13.2 An Example of Initial Misclassification

The liver scan data consist of images of 291 patients collected routinely between September 1976 and June 1977. Using brief extracts from the patient notes, relating both to the description of the scan and the clinician's opinion as to the patient's general condition, the scans were sorted into five classes.

1. Normal	71 cases
2. Cancer affecting the liver	46 cases
3. Diffuse liver disease	47 cases
4. Other diagnosed diseases	78 cases
5. Unsure	49 cases

The notes written by the doctors at the time of the scan were often vague; sometimes offering several possible diagnoses all of which would be consistent with the evidence then available. It would have been very useful had the doctors been asked to complete a simple questionaire at the time of the scan, stating their views and their degree of certainty. Unfortunately this was not done and by the time that this present analysis was undertaken it would not have been reasonable to ask the doctors to reassess the scans.

One is thus left with the feeling that, not only might

the diagnosis be wrong, but that the interpretation of the notes could also have introduced classification errors. Bearing all of these factors in mind it was decided to return to the patient records in order to try to improve on the quality of the classifications. This was carried out during the summer of 1982 and the results of the reassessment are summarised in table 13.1.

Table 13.1

Results of the reassessment

		SECOND ASSESSMENT						
		Nov	Can	Dif	Oth	Uns	N/T	Total
First Assessment	Nov	40	2	0	17	0	12	71
	Can	0	40	0	0	0	6	46
	Dif	10	1	22	9	1	4	47
	Oth	2	5	0	62	0	9	78
	Uns	5	4	2	11	14	13	49
	Total	57	52	24	99	15	44	291

Nor = normal Dif = diffuse Uns = Unsure

Can = Cancer Oth = Other N/T = not traced

The degree of disagreement is really quite marked; only 69% of those traced were in the same category at both assessments, what is more that percentage includes a number of 'other diseases' where the diagnosis changed but not the categorisation. Indeed the fact that 15% of cases could not be traced despite the possession of both the name and hospital record number is itself of concern.

Within the main, usable categories of normal, diffuse and cancer affected livers, we are left with only 102 cases with what might be considered a definite diagnosis. That is to say only 35% of the original data set, and of those five had to be discarded because of errors in the recording of the scans.

13.3 A Review of the Literature

The statistical and engineering papers which consider initial misclassification do so in almost total isolation from one another; there being very few cross references.

The first statistics paper to consider the problem was by Lachenbruch(1966). In that paper he considered the effect of random initial misclassification on the linear discriminant function. Thus suppose that the two classes w_1 and w_2 are multivariate normal with common covariance matrix $\underline{\Sigma}$ and means $\underline{\mu}_1$ and $\underline{\mu}_2$. Then the optimum linear classifier is,

$$D(\underline{x}) = (\underline{x} - \tfrac{1}{2}(\underline{\mu}_1 + \underline{\mu}_2))' \underline{\Sigma}^{-1} (\underline{\mu}_1 - \underline{\mu}_2)$$

Assuming however that a proportion a_1 of the n_1 cases classified initially as coming from w_1 actually come from w_2 and that a proportion a_2 of the n_2 classified as coming from w_2 actually come from w_1, then $D^*(x)$ the discriminant function estimated from the imperfect data will be, for large samples,

$$D^*(\underline{x}) = K\left[D(\underline{x}) - \frac{a_1 - a_2}{2} D^2\right]$$

where

$$D^2 = (\underline{\mu}_1 - \underline{\mu}_2)' \underline{\Sigma}^{-1} (\underline{\mu}_1 - \underline{\mu}_2)$$

and

$$k = \frac{1 - a_1 - a_2}{1 + \dfrac{a_1(1 - a_1)\, n_1 + a_2\,(1 - a_2)\, n_2}{n_1 + n_2}}$$

Futher the probabilities of subsequent misclassification become,

$$p_1 = P(\text{classify as } w_2 \,|\, \underline{x} \in w_1) = \Phi(-\tfrac{1}{2}D\,(1 + a_1 - a_2))$$

$$p_2 = P(\text{classify as } w_1 \,|\, \underline{x} \in w_2) = \Phi(-\tfrac{1}{2}D\,(1 + a_2 - a_1))$$

The consequences of these results are threefold,

(i) If $a_1 = a_2$ then the probabilities of subsequent misclassification are unaffected.

(ii) The estimate of D^2 based on the imperfectly classified data is actually estimating,

$$\frac{(1 - a_1 - a_2)^2\, D^2}{1 + k\, D^2}$$

that is to say it is underestimating the actual separation.

(iii) Since the estimate of the probability of subsequent misclassification depends on D^2, this will tend to be overestimated.

Lachenbruch goes on to consider the small sample case by means of a small scale simulation, concluding that the large sample results hold to a good approximation. McLachan(1972) obtained asymptotic solutions to the small sample equations and confirmed Lachenbruch's conclusions.

Lachenbruch(1974) returned to the problem of initial misclassification in order to see what effect the assumption of randomness had on the results. He conducted a series of small scale simulations in which all the cases more likely to have come from the other class were misclassified. The conclusions reached were very similar; that is,

(i) apparent error rates are seriously affected,

(ii) actual error rates are only slightly affected.

These generally reassuring findings do not extend to the quadratic discriminant function. Lachenbruch(1979) in yet another paper on the topic considered the problem of discriminating between two multivariate normal classes with different covariance structures. By means of simulations in which he introduced random initial misclassification he found that,

(i) Initial misclassification does seriously affect the ultimate error rate.

(ii) The effect becomes more severe as the differences between the covariance matrices increase.

(iii) The effect also becomes more marked as the initial misclassification rates increase.

The problem of imperfect training sets has also received spasmodic attention in the engineering literature. Kashyap(1970) described an optimisation procedure for dealing with the problem of two non-overlapping classes with initial misclassification at a known rate, and later Shanmugan and Breipol(1971) proposed a scheme based upon non-parametric density estimation. This method is essentially equivalent to estimating the likelihood ratio of each case and then reclassifying it if the ratio exceeds some pre-specified threshold. Various strategies for selecting the threshold were discussed but all required that the initial misclassification be random and at a known rate. The idea of a threshold was also used by Chittineni(1980,1981). He proposed an iterative scheme in which estimates of the misclassification rate are updated with the objective of minimising the eventual probability of misclassification.

The use of a threshold came up in another guise in a paper by Gimlin and Ferrell(1974). They proposed the use of a scheme based on the k nearest neighbours. In their scheme if more than k′ of the nearest neighbours are in the same class then the classification of that case is changed to that class. They suggested that the choice of metric and the choice of values for k and k′ are not critical and might be left to the subjective judgement of the investigator.

In a series of papers Chitti Babu(1972a, 1972b, 1973a,

1973b) considered the affects of imperfectly classified training sets on the process of feature selection. He showed that minimisation of the Bhattacharyya coefficient based on the imperfectly classified data minimises an upper bound on the eventual probability of misclassification. He then suggests that this justifies the use of the Bhattacharyya coefficient as the basis for feature selection when initial misclassification is suspected. By an identical argument he also shows that the minimisation of the divergence of imperfectly classified data minimises an upper bound on the eventual probability of error and that the error resulting from the use of a fixed number of terms from the Karhunen-Loeve expansion based on imperfectly classified data is bounded above by the error resulting from the use of the same number of terms together with correctly classified data.

In all cases the argument rests on the assumption that it is sensible to minimise a function that is bounded below by the function that we would actually like to minimise. He was not able to show that the minima occur close together or to indicate the tightness of the bounds and hence there must remain doubts over the usefulness of the results.

13.4 Iterative Discriminant Analysis allowing for initial misclassification

As we have seen several people have suggested the use of threshold values as a method of overcoming initial misclassification. Thus if, on the basis of the data, a case appears to have been much more likely to have come from a different class to that originally allocated then its classification is changed and the analysis is repeated. The process stopping when further allocation seems unnecessary.

One alternative to the threshold approach would be to allocate to each case a set of weights w_{ij}, where,

$i = 1, \ldots N$ denotes case

$j = 1, \ldots m$ denotes the class

$w_{ij} \geq 0$ and $\sum_{j=1}^{m} w_{ij} = 1$

Such weights could then be updated in the light of the analysis, once again the process being updated until convergence is obtained.

Clearly these weights have the characteristics of probabilities and there appears on the surface to be an analogy with the prior probabilities of a Bayesian analysis; this would imply that updating would be according to Bayes theorem using the likelihood of the value. The model is

however unlikely to be applicable in practice for the prior assessment of class membership is usually performed in full knowledge of the data. The problem is not that new evidence has come to light that neccesitates a change of opinion, but that a re-examination of the old evidence suggests that the judgement was wrong. Whilst Bayesian method show us how to update our views in the light of new information, they do not help when we wish to reassess our views in the light of a re-examination of the old data.

If we feel that misclassification is a problem then it would be better to have an analysis that whilst using the initial classification is not dependent on it for its final result. In practice, of course, there will usually be a mixture of some cases about which one is confident of the diagnosis and others about which one is less sure.

In the following sections we investigate the affect of iterative weighting on two of the more common methods of discrimination.

13.5 Iteratively Weighted kth Nearest neighbour

According to the kth nearest neighbour scheme one classifies a new case by estimating the probability that it comes from class w_j using k_j/k, where k_j is the number of cases out of the k nearest neighbours that originate from class w_j. Having found these probabilities the unclassified case is allocated to the class with the highest estimated probability. Essentially this procedure divides the measurement space S into mutually exclusive and exhaustive regions S_L. Each region is then associated with one of the possible classes with a strength that depends upon the estimated probability.

Suppose however that the training sets were suspected of containing misclassified cases. We might allocate an initial set of weights w_{ij} to each case then use a leave-one-out algorithm to re-estimate the weight on the basis of the other cases. Thus one would consider each case in turn, find its k nearest neighbours and make,

$$w_{ij} = \frac{\sum_L w^{o}_{Lj}}{K}$$

where L runs over the K-NN

This can then be made the basis for an iterative scheme whereby,

$$w_{ij}^{t+1} = \frac{\sum_{L} w_{Lj}^{t}}{K}$$

The result of such an iterative process can best be seen using the analogy with a Markov process. Suppose that we have five points A,B,C,D and E, as shown in figure 13.1, and that we are using k=2.

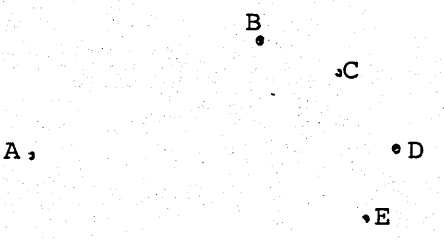


Figure 13.1

Five points to illustrate the Markov analogy

The updating of the weights will follow the matrix equation,

$$w_{ij}^{t+1} = \underline{\pi}\, \underline{w}_{ij}^{t}$$

$$\underline{\pi} = \begin{vmatrix} 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{vmatrix}$$

This is directly analogous to the matrix of transition probabilies for a five state Markov process, see for example Cox & Miller(1965). From this theory it can be shown that the weights will converge to,

$$\underline{w}_j = (\lim_{t\to\infty} \underline{\pi}^t)\, \underline{w}_j^o$$

Thus the process will divide the training data into sets of points which influence one another, corresponding to the analogous irreducible states of a markov chain. Thus,

$$\lim_{t\to\infty} \pi_{pq}^t = \frac{1}{u_q}$$

Where u_q is the mean recurrence time. Consequently the final weights will be,

$$w_{ij} = \sum_q \frac{w_{qj}^o}{u_q}$$

That is to say that within any irreducible set the points will all have the same weights.

Once again the final set of weights will divide the measurement space into regions within which the effect of any initial misclassification will have been averaged out.

As an example of the working of this iterative scheme consider the data displayed in figure 13.2. Here k=2 and a Euclidean measure of distance is being employed.

The data contain three deliberately misclassified values and in the first part of figure 13.2 we can see the dramatic effect that they have on the classification regions. The data will be seen to divide into three irreducible sets, namely all of class A, those cases in class B for which $x_2>0.1$, and those cases in class B for which $x_2<0.1$. Iterative analysis thus gives different weights to each of these regions. They are, in terms of the chance of coming from class A. 0.89, 0.28 and 0 respectively.

The resulting classification regions shown the different levels of certainty with which classification can be made, and almost entirely overcome the initial misclassification.

Clearly the benefit of this method would be lost were the data set not to reduce to a series of distinct sets. This problem is illustrated in figure 13.3 were there is such overlap between the classes that every point influences every other point, thus the iterative scheme would end up giving each point the same weight. What is required is that the data should be edited so that small independent sets of points are produced.

The data shown in figure 13.3 were randomly generated. Beside each point is the final weight allocated by an iterative 2-NN analysis in which a weight of 1 ia associated with class A and a weight of 0 is associated with class B. Even with this interacting data set there are two main groupings, but if further division is though desirable then

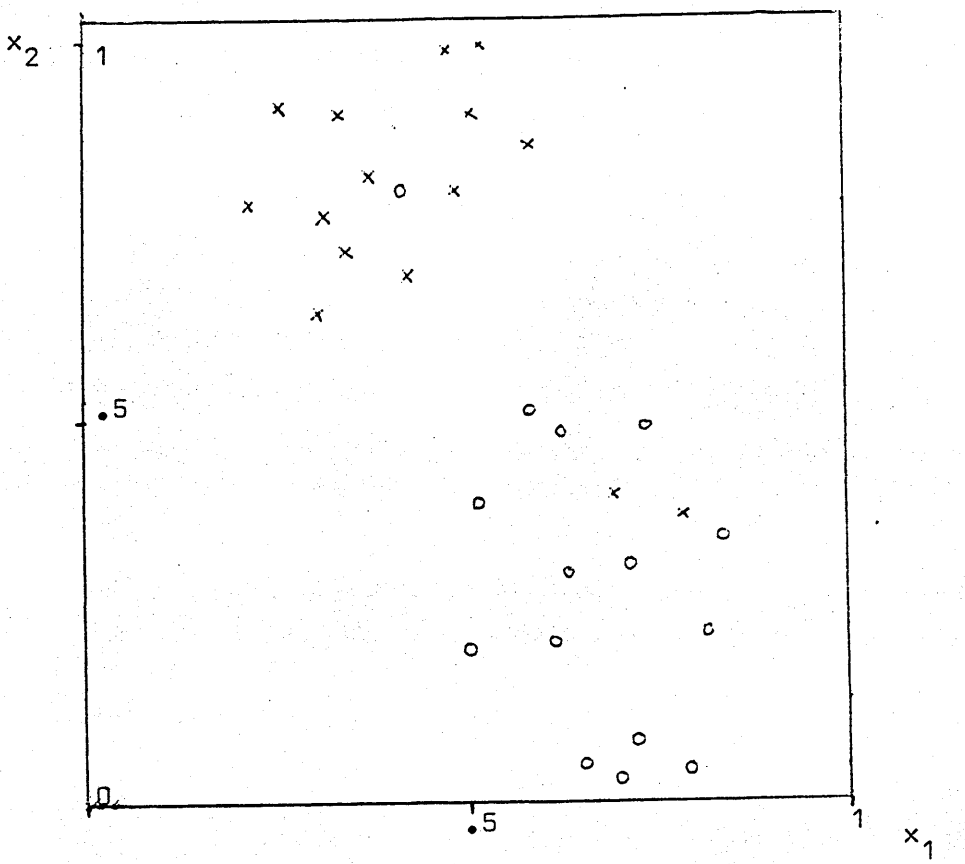


figure 13.2(a)

the data class A = x class B = o

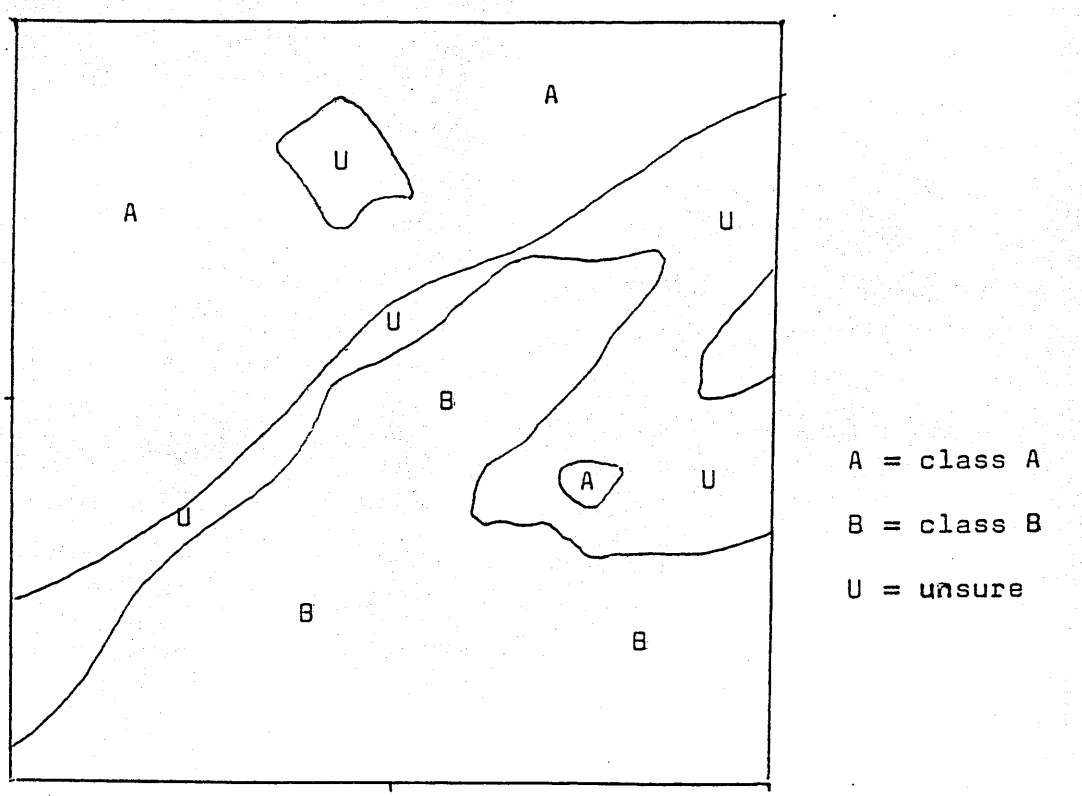


figure 13.2(b)

classification regions for the standard k-NN analysis

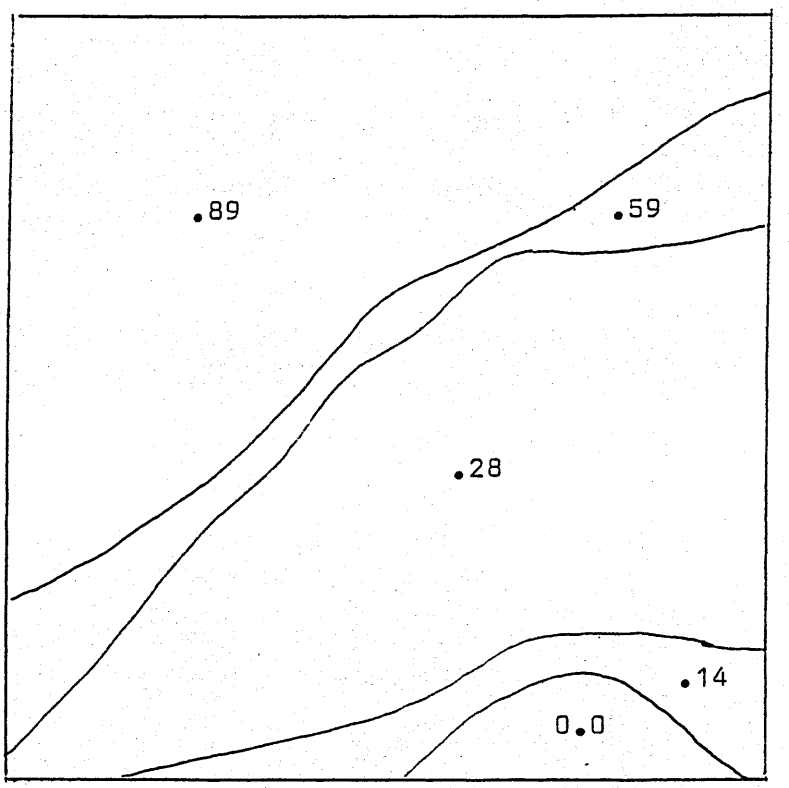


figure 13.2(c)

the results of the iterated analysis
showing the weight given to a point falling in each region
where the weight shows the chance that the value came from class A

some sort of editting is needed.

Editted nearest neighbour analysis has been quite widely studied, see for instance, Hart(1968), Wilson(1972), Koplowitz and Brown(1978) or Kittler and Devijver(1982). The methods of editing are based on the idea that one should seek to remove some of the data points in order to clarify the distinctions between the classes. This is generally not a good idea in an iterative analysis for removing a point merely forces the nearest neighbour link to go elsewhere and considerable forethought is necessary if one is to ensure that the situation will not be made worse by the new links.

The solution to this problem is to edit the links and not the data points. Thus in figure 13.4 we see the results of using only those 2-nn links that are less than one unit in length. The benefits are not great for this particular example but none the less they are clear. The cases in the bottom left and top of the diagram are isolated into separate groups and retain their original clear classifications.

In practice some experimentation and/or prior knowledge would be needed before specifying the threshold but this should not present a major difficulty.

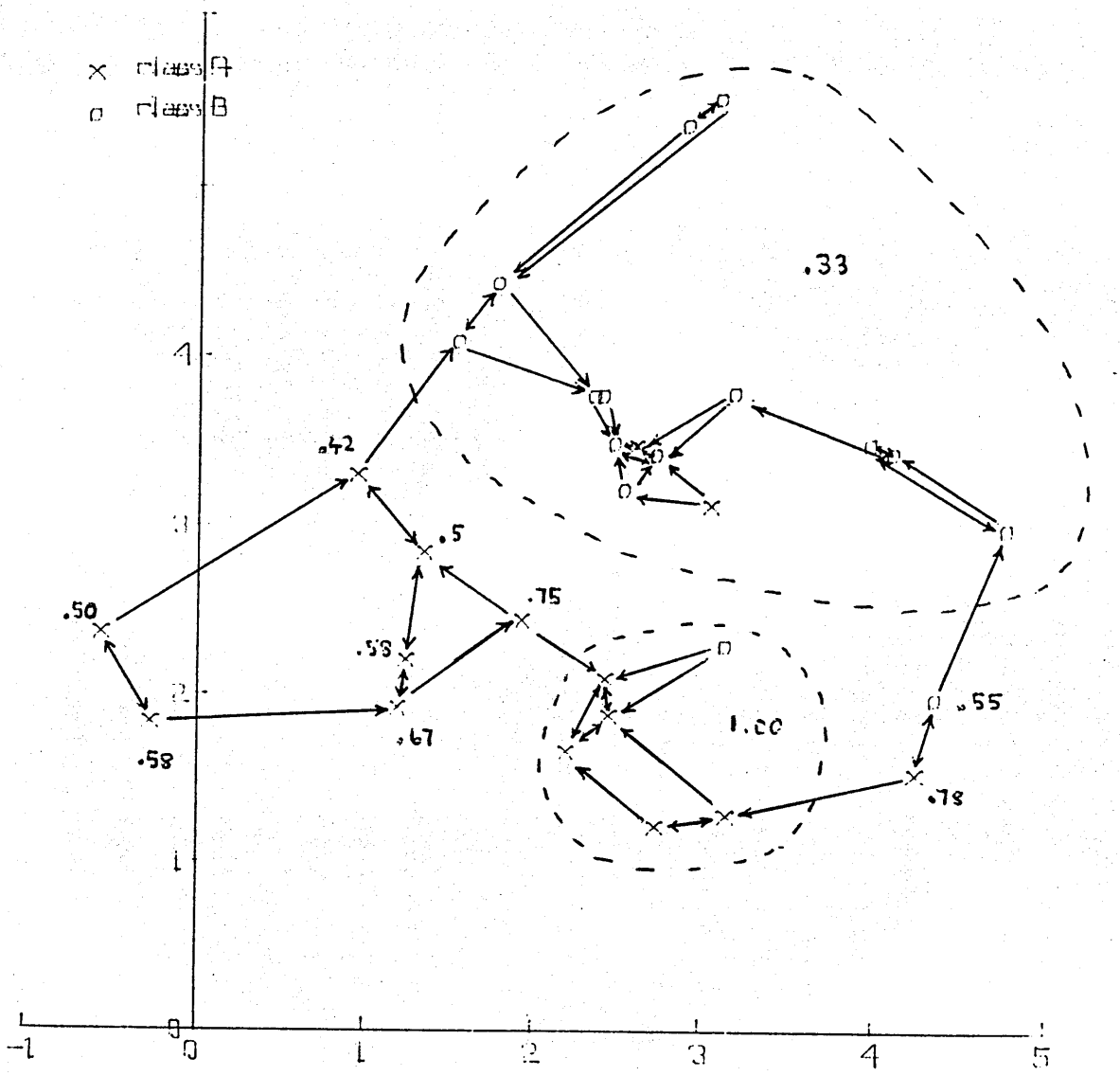


Figure 13.3

A Simulated Nearest Neighbour Analysis

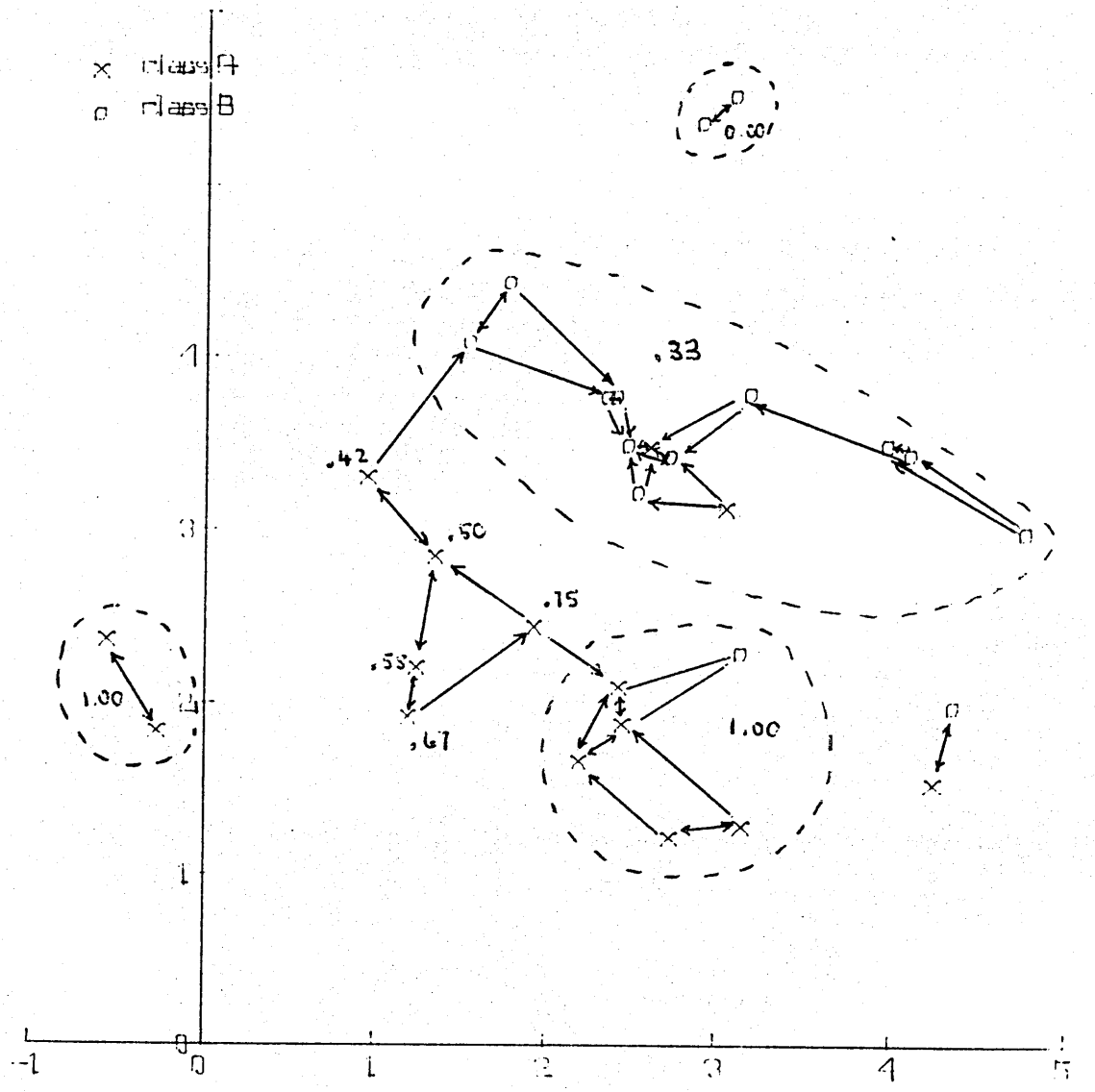


Figure 13.4

An iterative solution to the problem illustrated in Figure 13.3

13.6 Iteratively Weighted Linear Discrimination

For two normal distributions with a common covaraince matrix the discriminant function that minimises the probability of eventual misclassification is,

$$D(x) = (\underline{x} - \tfrac{1}{2}(\underline{\mu}_1 + \underline{\mu}_2)')\ \underline{\Sigma}^{-1}\ (\underline{\mu}_1 - \underline{\mu}_2)$$

This function can then be used together with the rule,

$$D(\underline{x}) \geq c \quad \text{assign to } w_1$$

$$< c \quad \text{assign to } w_2$$

where c is a constant depending on the prior probabilities and costs of misclassification.

Given two perfectly classified training sets containing n_1 and n_2 cases, it is usual to estimate the parameters by,

$$\hat{\underline{\mu}}_1 = \bar{\underline{x}}_i = \sum_{j=1}^{n_i} \underline{x}_{ij}/n_i$$

$$\hat{\underline{\Sigma}} = \underline{S} = \sum_{i=1}^{2} \sum_{j=1}^{h_i} (\underline{x}_{ij} - \bar{\underline{x}}_i)' (\underline{x}_{ij} - \bar{\underline{x}}_i)/(n_1 + n_2 - n)$$

and then to use the estimated linear discriminant function,

$$D_s(\underline{x}) = (\underline{x} - \tfrac{1}{2}(\bar{\underline{x}}_1 + \bar{\underline{x}}_2)')\ \underline{S}^{-1}\ (\bar{\underline{x}}_1 - \bar{\underline{x}}_2)$$

However, if misclassification is a problem then this estimate can be misleading, as Lachenbruch(1966) has shown.

The iteratively weighted algorithm would, in this case, consist of the following stages;

STAGE 1:

Set the initial weights reflecting the initial misclassification,

$$w_i^{o} = 1 \quad \text{if initially classified as } w_1$$

$$= 0 \quad \text{if initially classified as } w_2$$

STAGE 2: Form the weighted estimates, (initially with k=0),

$$\underline{\mu}_1^{k} = \bar{\underline{x}}_1^{k} = \left(\sum_{i=1}^{N} w_i^{k}\, \underline{x}_i\right) / \left(\sum_{i=1}^{N} w_i^{k}\right)$$

$$\underline{\mu}_2^{k} = \bar{\underline{x}}_2^{k} = \left(\sum_{i=1}^{N} (1 - w_i)^{k}\, \underline{x}_i\right) / \left(\sum_{i=1}^{N} (1 - w_i^{k})\right)$$

$$\hat{\underline{\Sigma}}^{k} = \underline{S}^{k} = \frac{\sum_{i=1}^{N} \{w_i^{k}\,(\underline{x}_i - \bar{\underline{x}}_i^{k})'(\underline{x}_i - \bar{\underline{x}}_1) + (1 - w_i^{k})(\underline{x}_i - \bar{\underline{x}}_2^{k})'(\underline{x}_i - \bar{\underline{x}}_2)\}}{N - 2}$$

STAGE 3: Use the relative likelihoods to define a new set of weights,

$$w_i^{k+1} = \frac{\exp\{-\tfrac{1}{2}(\underline{x}_i - \underline{x}_1^k)'\underline{S}^k(\underline{x}_i - \underline{x}_1^k)\}}{\exp\{-\tfrac{1}{2}(\underline{x}_i - \underline{x}_1^k)'\underline{S}^k(\underline{x}_i - \underline{x}_1^k)\}+\exp\{-\tfrac{1}{2}(\underline{x}_i - \underline{x}_2^k)'\underline{S}^k(\underline{x}_i - \underline{x}_2^k)\}}$$

STAGE 4:

If the weights have not converged,

i.e. $|w_i^{k+1} - w_i^k| > \varepsilon$ for some k

then return to stage 2.

If they have converged continue to stage 5.

STAGE 5:

Form the estimated linear discriminant function,

$$\underline{D}^*(\underline{x}) = (\underline{x} - \tfrac{1}{2}(\underline{\bar{x}}_1^k - \underline{\bar{x}}_1^k))'(\underline{S}^k)^{-1}(\underline{\bar{x}}_1^k + \underline{\bar{x}}_2^k)$$

An algorithm very similar to this has been extensively studied in the literature but not in the context of initial misclassification. It is, in fact, a method for solving the maximum likelihood equations for estimating the parameters of a mixture of two densities, namely,

$$L = \prod_{i=1}^{N} \{p(w_1)f_1(\underline{x}_i|w_1) + p(w_2)f_2(\underline{x}_i|w_2)\}$$

This approach to the estimation of the parameters of a mixture of two normals has been considered, amongst others, by Day(1969), Wolfe(1970), Homer(1973) and Murray and Titterington(1978). As Marriott(1975), Bryant and Williamson(1978) and McLachan(1980) have all found it is preferable to the reallocation of cases to their most likely class, a procedure that is effectively maximising the likelihood,

$$L = \prod_{i=1}^{N} f_1(\underline{x}_i|w_1)^{a_i} f_2(\underline{x}_1|w_2)^{1-a_i}$$

where a=0 or 1. This later method is found to produce biased estimates of the parameters.

Despite these objections reallocation continues to be a popular method of clustering, see for example, Friedman and Scott(1967), John(1970), Scott and Symmons(1971), Engelman and Hartigan(1969), Sclove(1977) and Everitt(1979).

The question as two how much is lost by not having classified observations when estimating the parameters has been studied by O'Neill(1978) and by Ganesalingam and McLachan(1978). O'Neil found that over the important range of separations (D=2.5 to 4), the amount of information in an unclassified case varied between 20% and 65% of that in a classified class. Ganesalingam and McLachan(1978) considered the problem in terms of the relative efficiencies of estimation based on classified and unclassified data.

They found that as the separation increased the efficiency of unclassified data approached two thirds of that of classified data.

Clearly then, if the choice is between classified data and unclassified data there will be a significant loss of information if one has to disregard all of the classifications. However there is also a price to be paid in terms of loss of performance if one uses data that was misclassified and these two factors need to be weighted against one another.

This is, of course, quite an extreme view, for in practice it is likely that the person performing the initial classification will be able to pick out those cases where the classification is uncertain. The likelihhod could then be modified to allow for n_1 and n_2 classified cases and n_u unsure ones, giving,

$$L = \prod_{i=1}^{n_1} f_1(\underline{x}_i|w_1) \prod_{i=1}^{n_2} f_2(\underline{x}_i|w_2) \prod_{i=1}^{n_u} \{p_1 f_1(\underline{x}_i|w_1)+p_2 f_2(x_i|w_2)\}$$

This modification has also been studied, in the context of parameter estimation, by Hartley and Rao(1968) and by McLachan(1975).

13.7 A Simulation Study of Iteratively Weighted Linear Discrimination

The use of an iteratively weighted linear discriminator entails a loss of information because one is effectively ignoring the initial classification; on the other hand, if initial misclassification is a problem then the failure to use an iteratively weighted analysis will also produce a loss in discriminatory power. In order to try to assess the relative importance of these two factors a simulation was performed.

The parameter combinations used are an extension of those suggested by Lachenbruch(1966) and consist of a factorial structure with,

dimension $p=2,4$ or 8

means $\underline{\mu}_1 = (0,0,\ldots,0)$

$\underline{\mu}_2 = (d,0,\ldots,0)$ $d=1,2,3,4$

sample sizes $n_1 = n_2 = 11,22$ or 44

proportions misclassified $a=0,0.091$ or 0.182

In each case we assess the resulting probability of error, averaged over one hundred simulations, contrasting the iteratively weighted analysis with the usual linear

discriminant function.

Various sampling methods are possible and in this study it has been assumed that the prior probabilities are known to be equal and that what are thought to be equal samples have been chosen from each of the two populations.

The results of the simulation are shown in tables 13.2, 13.3 and 13.4. For each of the parameter combinations the average resulting error rate and the standard deviation over the hundred simulations are both given. The iterative method converged in all cases although for large dimensions, large sample sizes and small separations the computation time was long.

Taking the two dimensional results from table 13.2 as an example, it will be seen that when the separation, delta, is either one or two, the initial error rate would need to be well in excess of our maximum of 18.2% before the iterative method would be justified. However as the separation increases to three so the 'breakeven' level of misclassification drops to between 9.1% and 18.2%. It will also be seen from the table of results that the larger the samples the smaller becomes the breakeven level of initial misclassification. When the separation is four then with all but the smaller sample size the iterative method is as good as the usual linear discriminant even when there are no initial errors.

Similar patterns emerge when one looks at the tables

showing the result for larger numbers of dimensions, although, not surprisingly, as the dimension increases so one needs either a larger sample size or a wider separation to achieve the same breakeven level of initial misclassification.

It might be noted in passing that as well as the usual linear discriminant function the simulation was also set to find the unbiased linear discriminant function. The resulting error rates are not shown as they were never significantly different from those shown in tables 13.2 to 13.4. However, what very small difference there was did always favour the unbiased version.

Simulation results for dimension p=2

delta	sample size	error rates a=0		a=.091		a=.182		Iterative	
1	11	.333	.030	.340	.034	.362	.046	.432	.073
	22	.320	.013	.331	.029	.355	.041	.404	.075
	44	.315	.007	.323	.009	.342	.016	.395	.067
2	11	.177	.020	.185	.027	.209	.043	.284	.124
	22	.166	.009	.172	.016	.191	.020	.242	.100
	44	.163	.005	.170	.010	.189	.015	.209	.072
3	11	.078	.012	.087	.023	.110	.034	.103	.051
	22	.075	.027	.079	.011	.097	.018	.086	.030
	44	.071	.023	.075	.005	.094	.010	.075	.011
4	11	.029	.007	.037	.017	.057	.028	.034	.017
	22	.029	.028	.032	.008	.046	.013	.028	.005
	44	.026	.024	.029	.003	.044	.007	.025	.002

Table 13.2

Simulation means and standard deviations

for dimension two

Simulation results for dimension p=4

delta	sample size	error rates a=0		a=0.091		a=.182		Iterative	
1	11	.366	.040	.372	.040	.378	.048	.427	.064
	22	.338	.026	.346	.031	.362	.032	.400	.057
	44	.323	.012	.325	.015	.329	.018	.362	.036
2	11	.205	.042	.218	.052	.240	.048	.254	.067
	22	.181	.030	.189	.021	.215	.031	.242	.057
	44	.170	.021	.176	.010	.194	.016	.229	.060
3	11	.099	.028	.115	.041	.143	.047	.151	.070
	22	.082	.033	.091	.017	.117	.027	.109	.036
	44	.076	.025	.089	.023	.099	.014	.089	.023
4	11	.037	.010	.057	.029	.081	.036	.058	.046
	22	.032	.026	.039	.013	.061	.021	.036	.018
	44	.029	.029	.033	.006	.051	.012	.028	.004

Table 13.3

Simulation Means and standard devaitions for dimension four

Simulation results for dimension p=8

delta	sample size	error rates a=0		a=.091		a=.182		Iterative	
1	11	.398	.038	.402	.043	.409	.042	.416	.043
	22	.362	.024	.370	.027	.378	.031	.399	.042
	44	.323	.012	.326	.015	.329	.018	.362	.036
2	11	.244	.048	.266	.056	.293	.061	.294	.068
	22	.198	.020	.214	.026	.242	.032	.270	.053
	44	.180	.011	.190	.016	.211	.020	.261	.049
3	11	.130	.038	.164	.053	.205	.062	.176	.068
	22	.098	.034	.113	.023	.142	.028	.161	.068
	44	.080	.006	.092	.011	.117	.019	.128	.045
4	11	.068	.031	.101	.047	.145	.064	.086	.051
	22	.042	.034	.057	.018	.089	.035	.059	.032
	44	.030	.004	.040	.008	.063	.016	.041	.019

Table 13.4

Simulation means and standard deviations
for dimension eight

14. AN ANALYSIS OF THE LIVER SCANS

14.1 The Features

The process by which the livers were characterised was described in chapter four. After a two dimensional spline surface had been fitted through the scan the residuals were calculated and these were then analysed by looking at their texture and for patches of negative residuals.

This characterisation lead to nine primary features which are both meaningful and which showed some potential for discrimination. They were,

1. The liver size
2. The average level over the liver
3. The standard deviation of the levels
4. The standard deviation of the residuals
5. The standard deviation of texture measurement one.
6. The standard deviation of texture measurement two.
7. The maximum level
8. The total size of the patches
9. The average depth of the residuals over the patches.

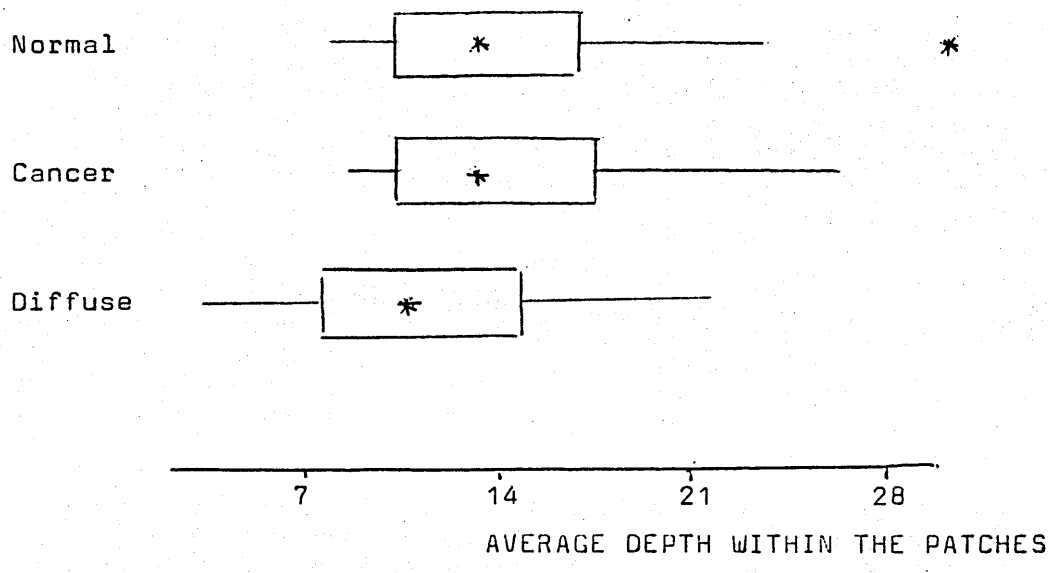
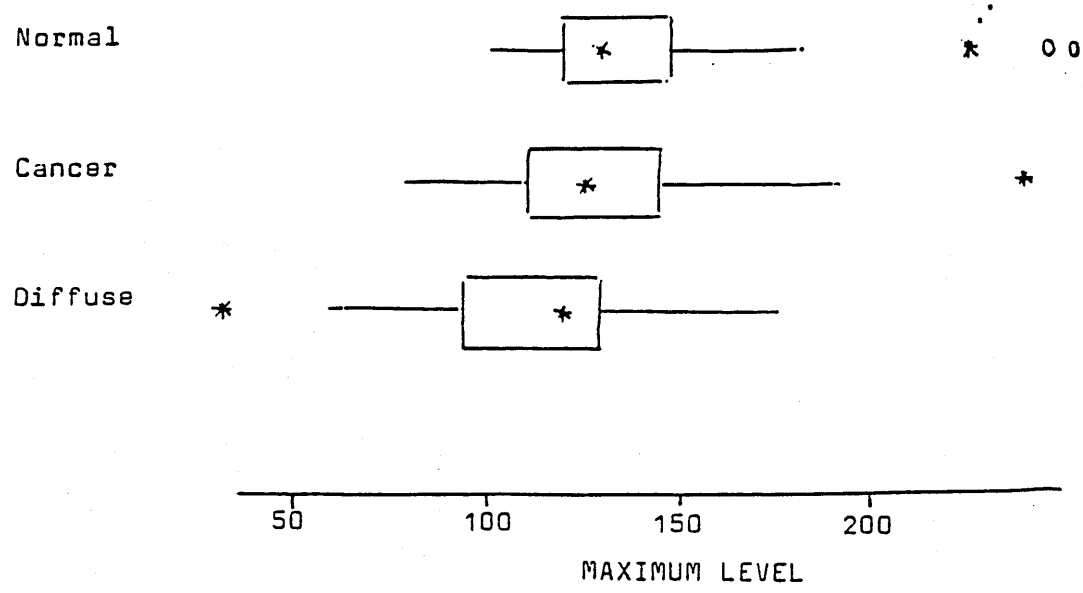
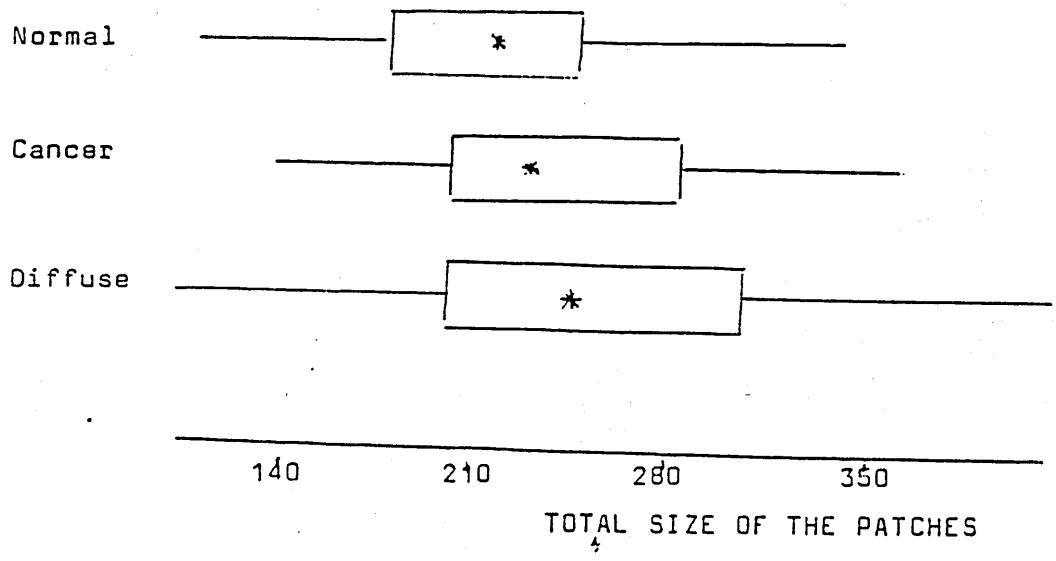
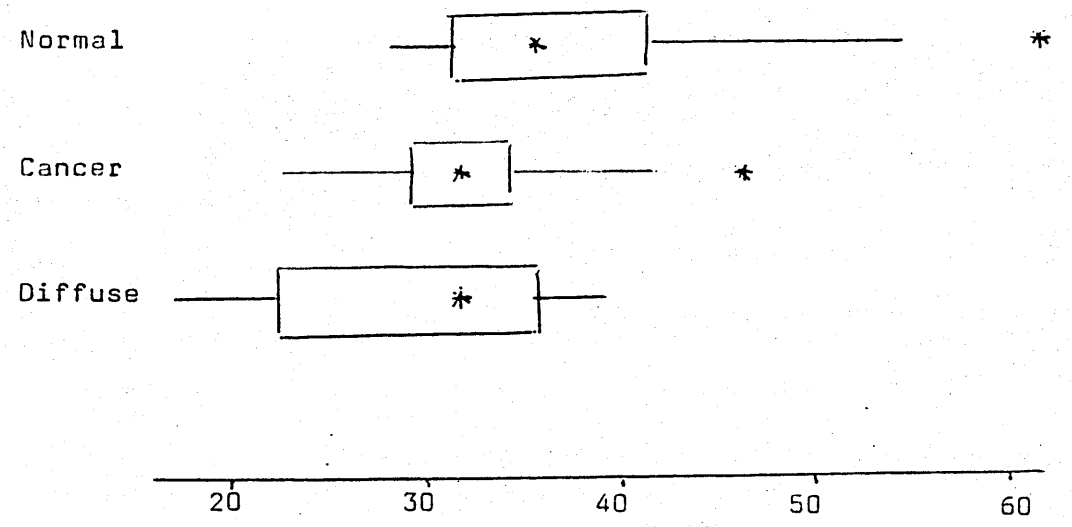


Figure 14.1

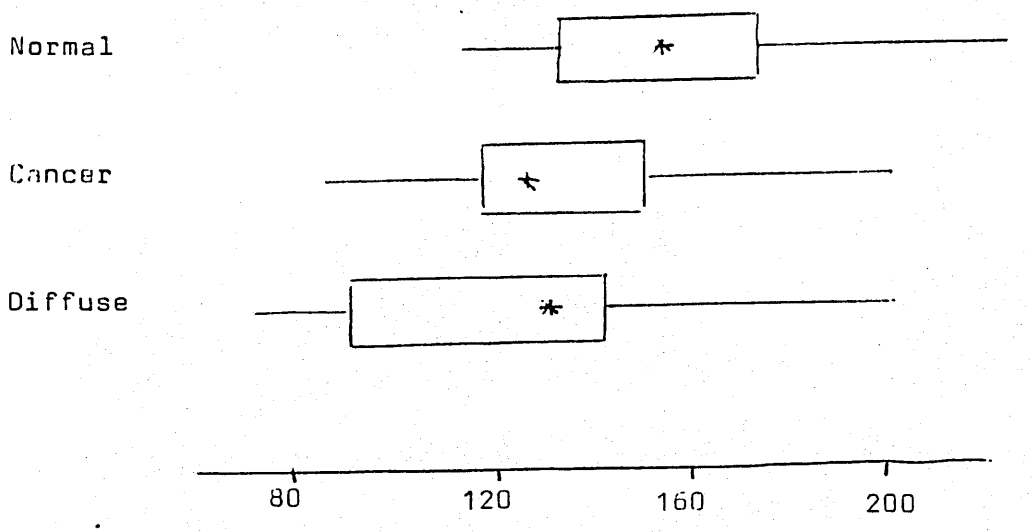
Boxplots of the nine primary features extracted from the liver scans


Normal
Cancer
Diffuse
50
100
150
200
MAXIMUM LEVEL

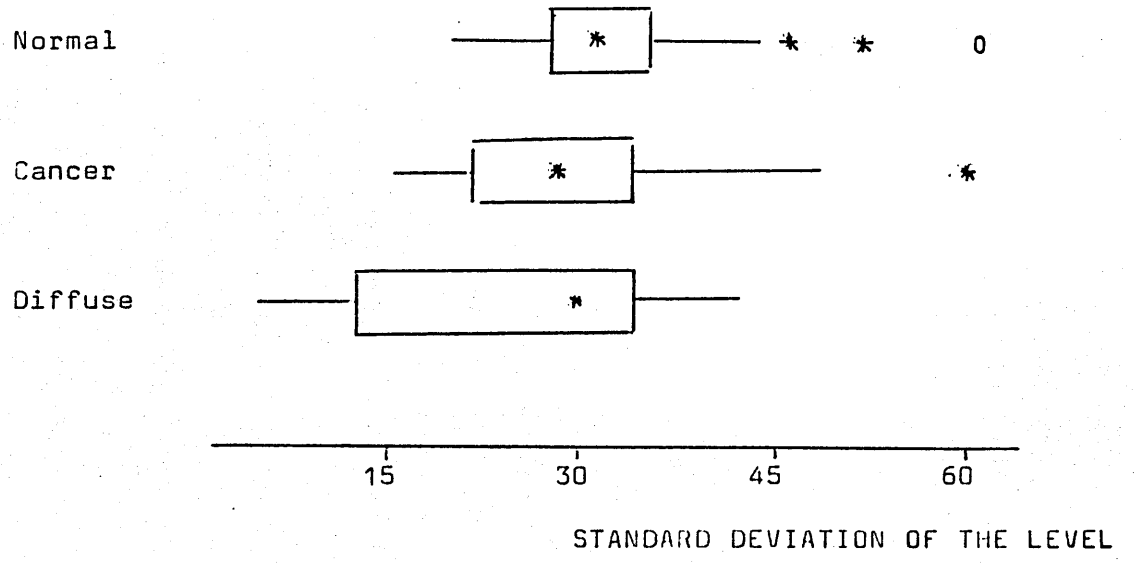

Normal
Cancer
Diffuse
140
210
280
350
TOTAL SIZE OF THE PATCHES

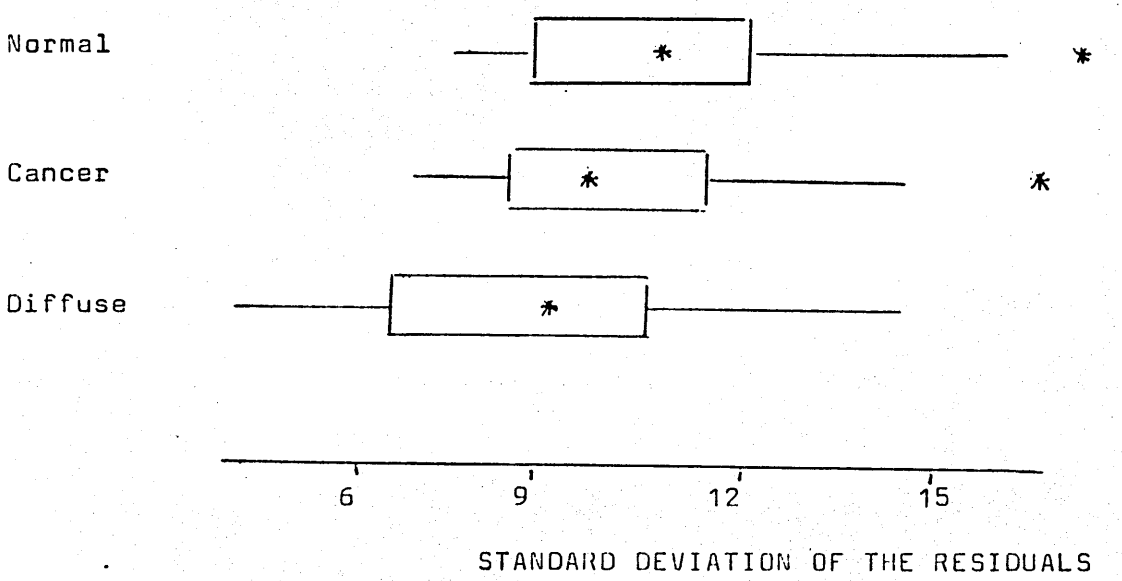


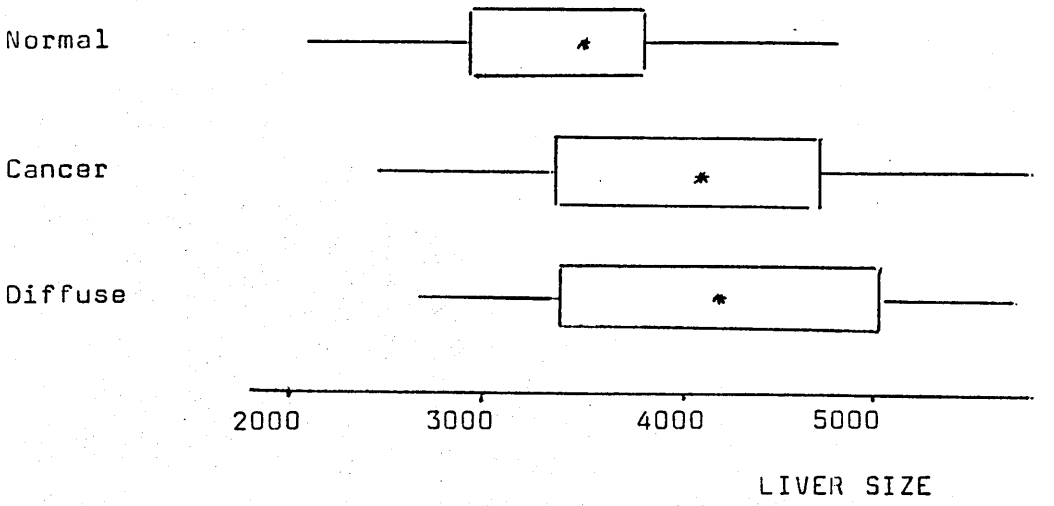
STANDARD DEVIATION OF TEXTURE MEASUREMENT ONE

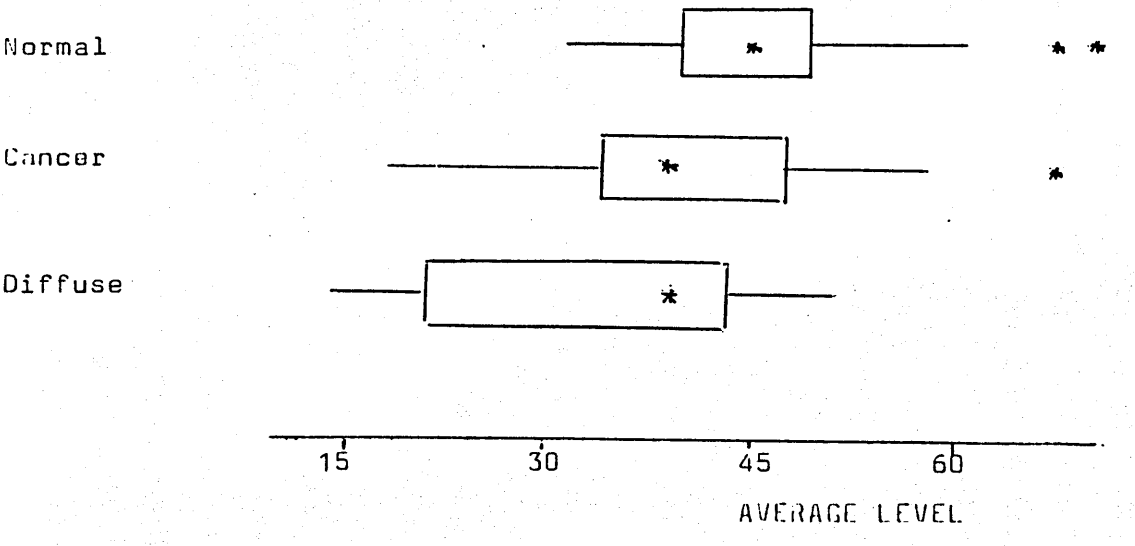


STANDARD DEVIATION OF TEXTURE MEASUREMENT TWO


Normal
Cancer
Diffuse
0
15
30
45
60
STANDARD DEVIATION OF THE LEVEL


Normal
Cancer
Diffuse
6
9
12
15
STANDARD DEVIATION OF THE RESIDUALS


Normal
Cancer
Diffuse
2000
3000
4000
5000
LIVER SIZE


Normal
Cancer
Diffuse
15
30
45
60
AVERAGE LEVEL

In figure 14.1 these nine features are boxplotted showing the distribution of the 39 normal cases, the 38 cancer affected cases and the 18 diffuse cases. Thus all of those cases for which the diagnosis was not consistent have been eliminated.

It will be noted that no feature gives a clear differentiation between the classes but that the normal livers tend to be smaller, have higher average uptakes, have larger texture variability and smaller patches than the abnormal livers. Generally speaking the diffuse livers tend to be further from the normals than do the cancer affected livers.

Figure 14.2 shows a biplot of the nine primary feaures and the 95 cases. Clearly there are two main groupings amongst the features measuring the size and level of uptake. The three classes are not well distinguished although there is a tendency for the diffuse cases to lie in the bottom left of the diagram and for the normal to lie if the top right with the cancer affected livers spread between. Indicating the relatively larger size and lower uptake of the diffuse cases.

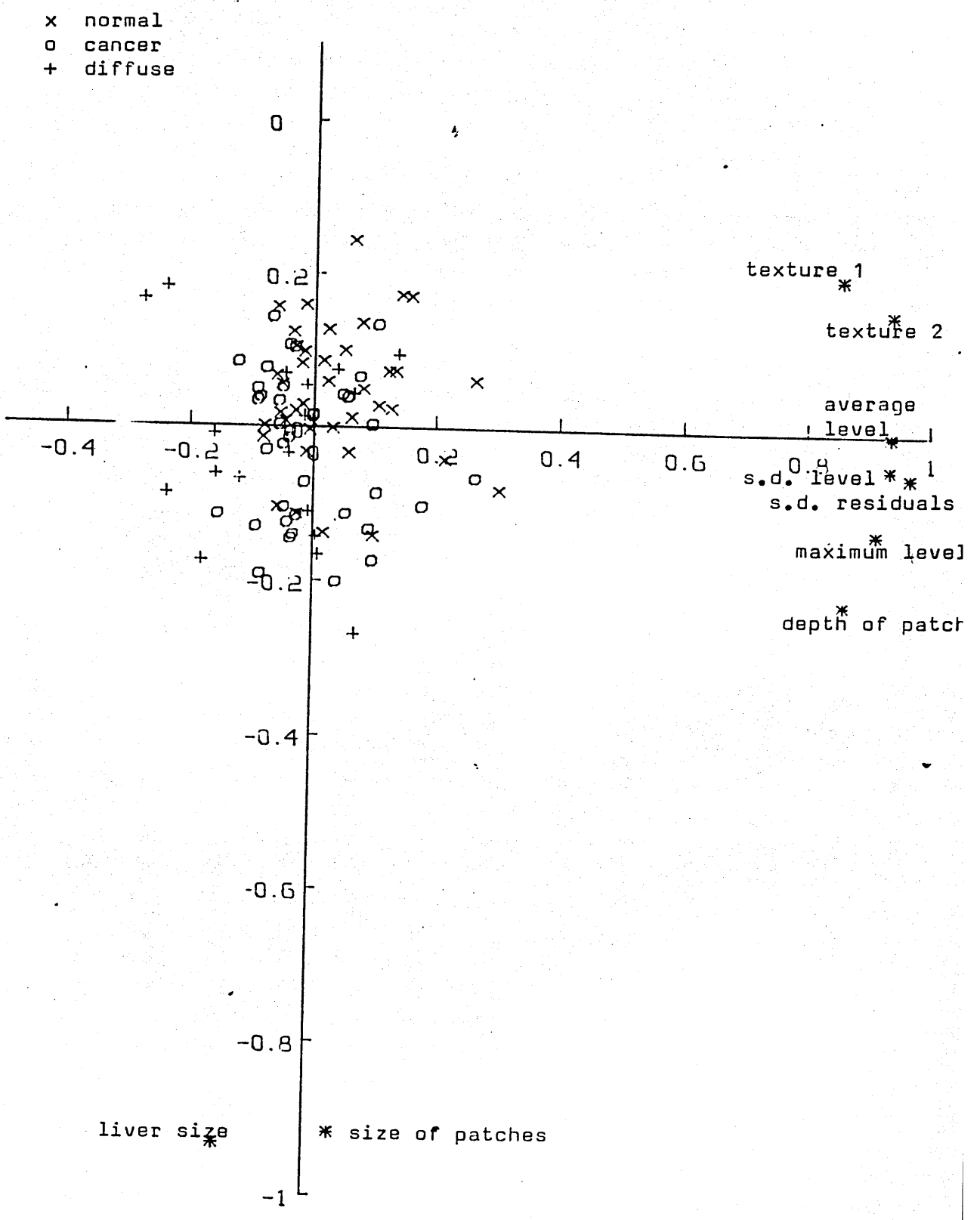


figure 14.2

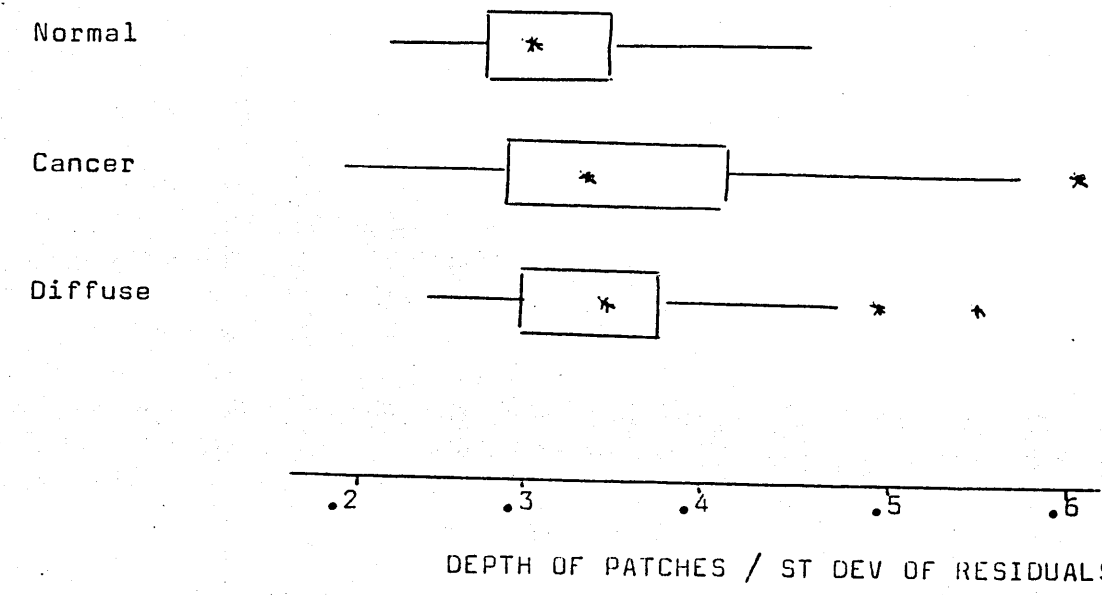
Biplot of the nine primary features

Four secondary features were formed from the nine primaries. In each case they were selected because of their natural interpretation. These were,

10. The total uptake, 1x2
11. The patches as a proportion of the liver, 8/1
12. The depth of the patches as a proportion of the standard deviation of the residuals, 9/4
13. The ratio of the standard deviations of the residuals to the levels, 4/3

These four secondary features are boxplotted in figure 14.3 from which it will be seen that features 11 and 13 appear to be the most useful when distinguishing between normal and abnormal livers. Feature 11 being primarily useful when detecting diffuse diseases.

The biplot of all thirteen features, figure 14.4, shows roughly the same distribution of cases but indicates that the new features do not merely duplicate the original nine. Feature 11 is rather poorly represented in the space of this diagram.



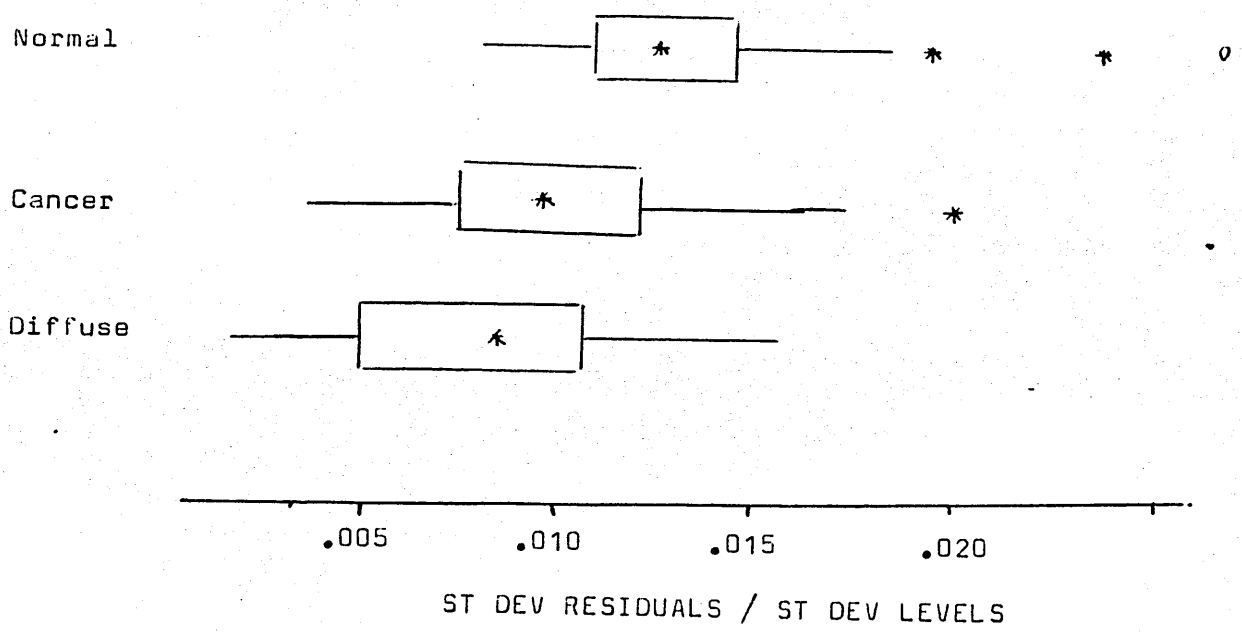
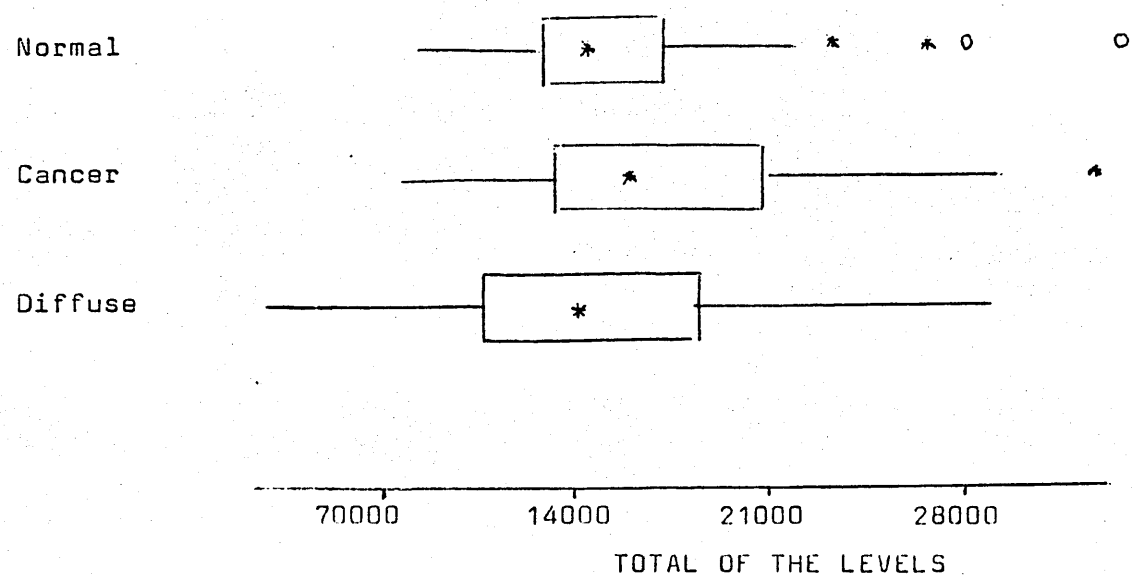
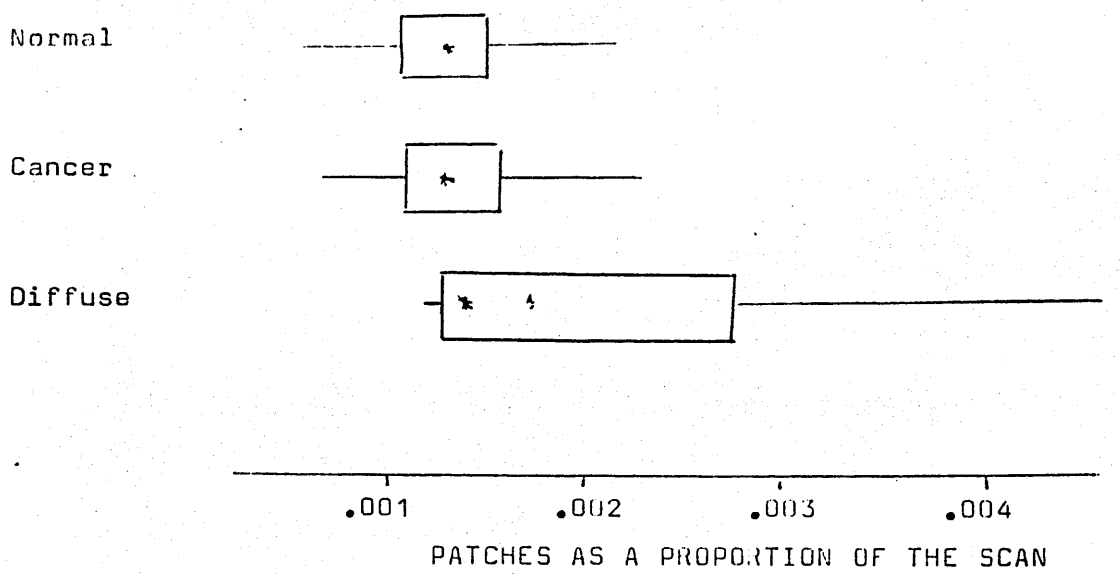


figure 14.3

Boxplots of the four secondary features


Normal
Cancer
Diffuse
70000
14000
21000
28000
TOTAL OF THE LEVELS


Normal
Cancer
Diffuse
.001
.002
.003
.004
PATCHES AS A PROPORTION OF THE SCAN

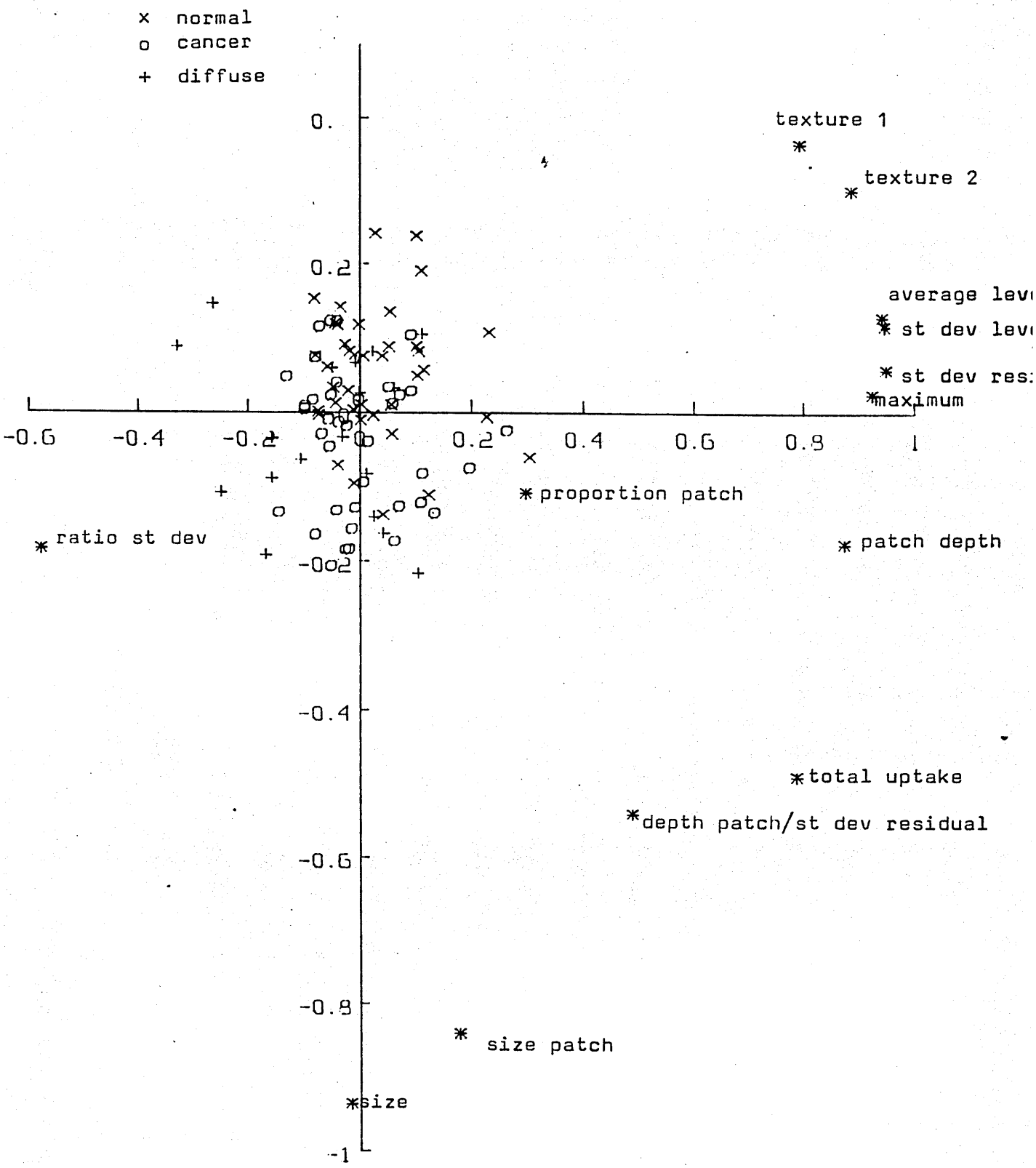


Figure 14.4

Biplot of the thirteen potential features

14.2 The Feature Selection

Four comparisons are of primary interest, they are the three possible comparisons between pairs of classes and the comparison between the normal class and the other two combined.

In each case the elimination method described in chapter six was employed using the proportion of the overlap as the selection criterion. The results are shown in tables 14.1 to 14.4. These tables show the first four features selected for each comparison, the values used when eliminating cases and the classification rates achieved by the sequential use of the selected features on the same data set.

The apparent success rates show clearly that it is only the normal vs diffuse case that is at all well differentiated; the other three all failing to reach an apparent success rate of 40%. These results reinforce the impression given by figures 14.1 to 14.4 that there is little hope of diffentiating the cancer affected livers from either of the other two classes.

Table 14.1

Feature selection - Normal vs Cancer

Feature	Separation	Classification normal	Classification Cancer	Apparent succes rate (cumulative
5. Texture one	-0.45	>47.9	<30.0	13%
1. Size	-0.59	<2600	>4920	22%
13. Ratio of st dev	-0.65	<.24	>.44	26%
9. Depth in patches	-0.81	>26.7	<9.3	29%

Table 14.2

Feature selection - Normal vs Diffuse

Feature	Separation	Classification normal	Classification diffuse	Apparent Succes rate(cumulative)
5. Texture one	-0.26	>41.7	<30.0	32%
13. Ratio st devs	-0.51	>.36	<.26	44%
9. Depth of patches	-0.51	<11.7	>18.2	67%
1. Size	-0.39	<3420	>4920	81%

Table 14.3

Feature selection - Cancer vs Diffuse

Feature		Separation	Classification Cancer	Diffuse	Apparent Success rate (cumulative)
7.	Maximum	-0.45	>180	<85	13%
8.	Patch size	-0.65	<180	>360	23%
10.	Total uptake	-0.65	>244 000	<110 000	29%
6.	Texture two	-0.53	<97.8	>157.8	36%

Table 14.4

Feature selection - Normal vs Rest

Feature		Separation	Classification Normal	Rest	Apparent Success rate (cumulative)
13.	Ratio st dev	-0.37	<.24	>.44	12%
5.	Texture one	-0.40	>47.9	<30.0	23%
1.	Size	-0.61	<2600	>4900	32%
9.	Depth Patches	-0.81	>26.7	<9.3	34%

14.3 Normal Livers vs Diffuse diseases

Using the four selected variables one might seek a linear combination to distinguish between normal and diffuse cases based on a single score. The methods employed on the psychiatric data in chapter twelve, with p=0.1, result in an overlap of 19.5%, as shown in figure 14.5.

The analysis is clearly not good enough for practical implimentation, giving as it does an apparent error rate of 39%. Changing the value of p to guard against outliers and misclassified cases does not enhance the overall performance.

Normal is Diffuse

Separation R = -0.192

Feature	Weight
1. Size (x.01)	.055
5. Texture one	-.648
9. Depth of patches	-.318
13. Ratio of st. dev(x 100)	.690

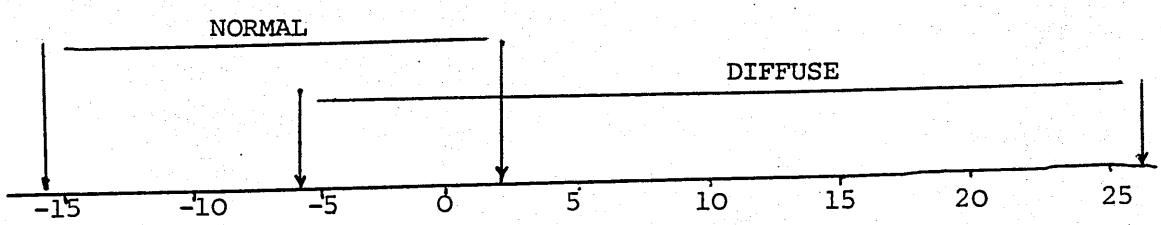


Figure 14.5

The separation between the normal and diffuse cases

14.4 Post Mortem

The analysis of the liver scans has not provided us with a practical scheme for the diagnosis of liver disease and it is important that we question the reasons for this failure. Several major factors can be identified as contributing to the difficulties, of these the problem of diagnosis has already been discussed in chapter thirteen.

A further difficulty is that the normal cases are not truely 'normal'. The data were routinely collected from subjects suspected as having a liver disease, the normal group thus contains those who turned out not to have diseased livers even though they had some symptoms that initially suggested that possibility. Whilst it is important to be able to distinguish such patients from those who actually have liver disease this is a far more difficult problem than that of distinguishing between liver disease and healthy livers. The simpler problem would be better suited to an initial investigation of this type.

Although the data set appears at first sight to be large we have seen that the way in which it was collected, without specific diagnosis has greatly reduced the effective sample size. The methods for dealing with initial

misclassification, described in chapter thirteen, might have been appropriate had the classes been better separated and had the important features been easier to identify.

The classes used in the analysis are not even homogeneous. Because of the uncontrolled nature of the initial study there are insufficient cases to make up classes of patients with a single disease. Thus diffuse disease includes, amongst others, cirrhosis and hepatitis, and even hepatitis is itself a collection of different forms. Clearly there is the strong possibility that methods suitable for picking out a case of cirrhosis may not be suitable for hepatitis and vice versa.

It has already been pointed out how the data was collected without sufficient background information on the patients. In the original study there was no record of the patients height, weight or age, factors that would have a influence on ones interpretation of the scan. When the patient records were traced in the summer of 1982 it was found that this information was only available on a fraction of the patients and to require a full set of patient details would have virtually eliminated all of the cases. Clearly this information should have been sort at the time of collection.

The method for detecting patches of negative residuals, indicating the presence of secondary cancers has not been effective. Had the data set been more reliable there are

many ways in which the method could have been modified in order to try to improve its performance. Chief amongst the variations would be an increase in the number of knots in the spline fit and the use of a robust fitting method. Spline fitting is based on least squares and it would be quite possible to weight the least square fit so as to play down the effect of the negative residuals, as illustrated in figure 14.5. Some experiments were made along these lines but the poor quality of the data made their continuation a doubtful exercise.

15. CONCLUDING REMARKS

At the outset the intention behind this work was that it should apply statistical pattern recognition to a substantial data set and that the application should motivate whatever theoretical developments that happened to prove necessary. Unfortunately two of the data sets proved to be a little disappointing. The ballistocardiograms because they were repeat measurements on so few subjects, and the liver scans because of the uncertainty over their correct classification.

Although the data sets were themselves not ideal they did still serve to motivate some developments that seem worth considering for future application.

The double damped harmonic model described in chapter four gives such a good fit to the ballistocardiograms that the parameters of the model allow one, not just to identify diseased cases, but even to identify the individuals. This model would make a useful basis for a further study, especially as the parameters have clear interpretations in terms of amplitude, phase and damping. The requirement now is for a larger data set with detailed covariate information such as sex, age and diagnosis. The model fitting could then be programmed into a microcomputer and used by doctors.

Because of the errors in the data base it is much more difficult to be confident as to the success or otherwise of the B-spline model for the liver scans. The degree of fit

does however suggest that they give a good representation of the overall pattern of uptake. It remains to be seen whether or not this representation leads to useful discriminatory features.

The liver scans did, none the less, motivate a small study of initial misclassification which suggested the use of the iterative maximum likelihood methods with misclassified data. Futher the iterative idea is extended to encompass the commonly used technique known as nearest neighbour analysis so that it can also be modified to allow for any initial misclassification.

The work on feature selection has potentially far reaching application and divides neatly into two parts, as described in chapters six and seven.

The conditioning method seems to be a useful way of selecting variables in problems that are large in terms of dimension and sample size. Whilst not optimal, the method works well in practice, is simple to program and quite general. Indeed, as has been shown, it can be applied both to assumed distributional forms and in non-parametric analyses. There is further scope for extension here by using the conditioning algorithm in conjuction with non-parametric density estimates.

Kernel estimates offer a good practical method for handling problems where the form of the density is unlikely to be closely approximated by a multivariate normal. Unfortunately kernel estimates are difficult to obtain in

high dimensions because of the large amounts of data that they require. The conditioning algorithm offers a useful way of reducing the selection to a series of univariate problems, essentially by not trying to estimate the density over regions that are of no interest.

The problems resulting from the use of this technique have by no means all been solved. Further work is required on the effect of the conditioning range on the forms of the resulting densities. As section 6.5 shows, for multivariate normal data, the normality of the variables is lost once they have been conditioned on a range of unhelpful values. However the examples investigated suggest that the effect of the non-normality is small and of little importance provided that a robust criterion has been used for feature selection.

The technique described in chapter six is sequential and univariate and yet tackles a problem that is multivariate. Clearly the results will not be optimal, although the method has the advantage of being applicable even when the full multivariate solution is prohibitively complex. It would be useful to have more information on the performance of the method in comparison with the optimal procedures, where these exist. The simulations reported in chapter six suggest that the technique performs very well, at least with normal data, but further work is needed to confirm this.

In chapter five the standard methods of feature selection were reviewed and it was shown how many depend on the

eigenstructure of some covariance estimate. In chapter seven it is argued that this idea is basically mistaken and that it is far better to look at the eigenstructure in the dual space. One again the method is far reaching and applicable to almost any problem. Possible objections due to the requirement of the eigenstructure of an nxn matrix, where n is the number of cases, are overcome by clustering the cases, using the cluster centres in the analysis and then replacing the individuals in the reduced space.

The work on non-parametric variable selection lead to the use of the non-parametric measure of separation based on tolerance intervals. This work forms a natural extension to the earlier ideas of Quesenbury and Gessaman(1968) selecting the linear function that gives the greatest separation. The overall method works well on both the ballistocardiograms and the psychiatric data.

Extension into higher dimension links further with a suggestion made, amonst others, by Kendall(1966), for the use of convex hulls in discrimination. In order to facilitate this extension an algorithm was developed, the first to allow the calculation of a convex hull in any number of dimensions. This algorithm may well find application in other areas.

In consequence, although the applications have proved disappointing, the ideas that they have generated seem to have considerable potential use.

REFERENCES

Aitchison J. & Dunsmore I.R. (1975)
Statistical prediction analysis
Cambridge University Press

Aichison J., Habbema J.D.F., & Kay J.W. (1977)
A critical comparison of two methods of statistical discrimination
Applied Statistics 26 pp15-25

Anderberg M.R. (1973)
Cluster Analysis for Applications
Academic Press
New York

Anderson J.A. (1972)
Separate sample logistic discrimination
Biometrika 59 pp19-35

Anderson J.A. & Richardson S.C. (1979)
Logistic discrimination and bias correction in maximum likelihood estimation
Technometrics 21 pp71-78

Anderson T.W. (1958)
An introduction to multivariate analysis
Wiley
New York

Anderson T.W. (1973)
An asymptotic expansion of the distribution of the 'studentised' classification statistic W
Annals Statistics 1 pp964-972

Anderson T.W. & Badahur R.R. (1962)
Classification into two multivariate normal distributions with different covariance matrices
Annals Math. Statist. 33 pp 420-431

Andrews H.C. (1972)
Introduction to mathematical techniques in pattern recognition
Wiley
New York

Angehrn F.G., Schmid P., Percia R., Cueni B., Schmid M.,
Akovbiantz A., Heinzel F., Landoft M., Haemmerli U.P. & Baum A.L.(1
Lebermetastasen: diagnostischer wert von bluttests,
szintigraphic und laparoskopic
Dtsch Med Wochenschr 101 pp1047-1055

Avis D. (1979)
On the complexity of finding the convex hull of a set of points
Technical report No. SOCS 79.2
School of Computer science
McGill University

Backer & De Shipper (1977)
On the max-min approach for feature ordering and selection
Proc. seminar on pattern recognition Liege

Bahadur R.R. (1961)
A representation of the joint distribution of response to
n dichotomous items
in
Studies in item analysis & prediction
Ed. Solomon H.
Stanford University pr Press
Palo Alto, California

Bard Y. (1974)
Non-linear parameter estimation
Academic Press London

Barnett V. (1976)
The ordering of multivariate data
J.R.S.S. A 139 pp318-354

Barr D.R. & Slezak N.L. (1972)
A comparison of multivariate normal generators
CACM 15 pp1048-1049

Batchelor B.G. (1974)
Practical approach to pattern classification
Plenum
New York

Becker P.W. (1971)
An introduction to the design of pattern recognition
devices
Springer
New York

Ben-Bassat M. (1980)
Multimembership and multiperspective classification: Introduction, applications and a Bayesian model
IEEE Trans Systems, Man & Cybernetics SMC-10 pp331-336

Bentley J.L. & Shamos M.I. (1977)
Divide and conquer for expected linear time
Carnegie-Mellon
University Computer Science Technical Report

Berkson J.T. (1955)
Maximum likelihood and minimum chi-squared estimation of the logistic function
J.A.S.A. 50 pp130-162

Biello D.R., Levitt R.G., Barry A.S., Sagel S.S. & Stanley R.J. (1
Computer Tomography and radionuclide imagary of the liver: a comparitive evaluation
Radiology 127 pp159-163

Blahd W.H. (1971)
Nuclear Medicine
McGraw-Hill New York

Blashfield R.K. & Aldenderfeld M.S. (1978)
The literature on cluster analysis
Multivariate behaviour research 13 pp271-295

Bowker A.H. (1947)
tables of tolerance factors for normal distributions
in
techniques of statistical analysis
McGraw Hill
new York

Box G.E.P. & Jenkins G.M. (1976)
Time series analysis
Holden-Day
San Francisco

Braunstein P. & Song C.S. (1975)
The uses and limitations of radioisotopes in the investiaggation of gastrointestinal diseases
Digest. Dis. 20 pp53-90

Brent R.P. (1974)
Algorithm 488: A Gaussian pseudo-random number generator
Communications of the ACM 17 pp704-705

Broffit , Randles , & Hogg (1978)
Discriminant analysis based on ranks
J.A.S.A. 73 pp379-384

Brofitt J.D., Rundles R.H. & Hogg R.V. (1976)
Distribution free partial discriminant analysis
J. Amer. Statist. Assoc. 71 pp 934-939

Bryant P. & Williamson J.A. (1978)
Asymptotic behaviour of classification maximum likelihood estimates
Biometrika 65 pp273-281

Bykat A. (1978)
The convex hull of a finite set of points in two dimensions.
Inf Process Lett 7 pp296-298

Caetano D. (1980)
Enquiries into the classification of affective disorders
Unpublished PhD dissertation
University of Cambridge

Chaud D.R. & Kapar S.S>. (1970)
An algorithm for convex polytopes
Journal of the ACM 17 pp78-86

Chen C.H. (1973)
Statistical pattern recognition
Hayden
New York

Chien Y.T. (1978)
Interactive pattern recognition
Marcel Dekker
New York

Chein Y.T. & Fu K.S. (1967)
On the generalised Karhunen-Loeve expansion
IEEE Trans. IT-15 p 518

Chitti Babu (1972a)
On the application of divergence for the extraction of features from imperfectly labeled patterns
IEEE trans System, Man & Cybernetics pp290-292

Chiitti Babu (1972b)
On the extraction of pattern features from imperfectly identified samples
IEEE trans computers pp410-411

Chitti Babu (1973a)
On the extraction of imperfectly labeled patterns
IEEE trans Systems, Man & Cybernetics pp290-292

Chitti Babu (1973b)
On the application of probabilistic distance measures for the extraction of features from imperfectly labeled patterns
Int J Comp & Inform Sc pp103-114

Chittineni (1980)
Learning with imperfectly labeled patterns
Pattern recognition pp281-291

Chittineni (1981)
Estimation of the probabilities of label imperfections and correction of mislabels
Pattern Recognition pp257-268

Clunies-Ross C.W. & Riffenburgh R.H. (1960)
Geometry and linear discrimination
Biometrika 47 pp185-189

Cochran W.G. & Hopkins C.E. (1961)
Some classification problems with multivariate qualitative data
Biometrics 17 pp10-32

Conn H.O. & Elkington S.G. (1968)
Is hepatic scanning overrated?
Gastroenterology 54 pp135-140

Cormack R.M. (1971)
A review of classification
J.R.S.S. A 134 pp321-367

Costanza M.C. & Afifi A.A. (1979)
Comparison of stopping rules in forward stepwise discriminant analysis
J. Amer. stat. assoc. 74 pp 777-785

Cover T.M. (1974)
The best two independent measurements are not the two best
IEEE Trans. syst. man cyber. (corresp) SMC-4 pp 116-117

Cover T.M. & Hart P.E. (1967)
Nearest neighbour pattern classification
IEEE Trans. Inform. Theory IT-13 pp 21-27

Cox D.R. (1970)
The analysis of binary data
Methuen
London

Cox D.R. & Miller H.D. (1965)
The theory of stochastic processes
Methuen
London

Cunningham D.M. & Smiley P.C. (1961)
J App Physiology 16 p4

David F.N. & Johnson N.L. (1954)
Statistical treatment of censored data
Biometrika pp228-240

Day N.E. (1969)
Estimating the components of a mixture of two normal distributions
Biometrika 56 pp463-473

Day N.E. & Kerridge D.F. (1967)
A general maximum likelihood discriminant
Biometrics 23 pp313-323

Devijver P.A. (1979)
New error bounds with the nearest neighbour rule
IEEE Trans. Inform. Theory IT-25 pp 749-753

Devijver P.A. & Kittler J. (1982)
Pattern recognition: A statistical approach
Prentice-Hall

Duda R.V. & Hart P.E. (1973)
Pattern classification & scene analysis
Wiley
New York

Eckart C. & Young G. (1939)
A principal axis transformation for non-Hermitian matrices
Ann Math Soc Bull 45 pp118-121

Eddy W.F. (1977)
A new convex hull algorithm for planar sets
ACM trans on Math Software 3 pp398-404

Efron B. & Morris C. (1976)
Multivariate empirical Bayes and estimation of covariance matrices
Annals Statist. 4 pp 22-32

Eisenbeis R.A. & Gilbert G.G. (1973)
Investigating the relative importance of individual variables and variable subsets in discriminant analysis
Commun. in stats. 2 pp 205-219

El-Sheikh T.S & Wacker A.G. (1980)
Effect of dimensionality and estimation on the performance of Gaussian classifiers
Pattern Recognition 12 pp115-126

Engelman L. & Hartigan J.A. (1969)
Percentage points of a test for clusters
J.A.S.A. 64 pp1647-1648

Everitt B.S. (1979)
Unsolved problems in cluster analysis
Biometrics 35 pp169-181

Everitt B.S. (1980)
Cluster analysis
Heinemann
London

Fehlauer J. & Eisenstein B.A. (1979)
A declustering criterion for feature selection in pattern recognition
IEEE trans computers C-27 pp261-265

Fisher R.A. (1936)
The Use of multiple measurements in taxonomic problems
Ann Eugen 7 pp 179-188

Fix F. & Hodges J.L. (1951)
Discriminatory analysis: non-parametric discrimination: small sample performance
US School of Aviation Medicine Proj 21-49-004
Rep 11 Randolph Field Texas

Foley (1973)
Orthogonal expansion study for waveform processing systems
Rome air develop. centre AF systems command
Griffiss AEB New York Tech. Rep. RADC-TR-73-168

Friedman J.H. & Tukey J.W. (1974)
A projection pursuit algorithm for exploratory data analysis
IEEE Trans Comp C-23 pp881-889

Friedman H.P. & Rubin J. (1967)
On some invariant criteria for grouping data
J.A.S.A. 62 pp1152-1178

Fu K.S. (1968)
Sequential methods in pattern recognition and machine learning
Academic
New York

Fu K.S. (1974)
Syntatic methods in pattern recognition
Academic
New York

Fu K.S. (1980)
Recent developments in pattern recognition
IEEE trans on computers C-29 pp845-854

Fukunaga K. (1972)
Introduction to statistical pattern recognition
Academic Press New York

Fukunaga K. & Kessel D.L. (1971)
Estimation of classification error
IEEE trans on computers C-20 pp1521-1527

Fukunaga K. & Koontz (1970)
Application of the Karhunen-Loeve expansion to feature selection and ordering
IEEE Trans. Comp. IC-19 p 311

Gabriel K.R. & Zamir S. (1979)
Lower rank approximation of matrices by least squares with any choice of weights
Technometrics 21 pp489-498

Ganesalingam G. & McLachan G.J. (1978)
The efficiency of a linear discriminant function based on unclassified initial samples.
Biometrika 63 pp658-662

Geisser S. (1964)
Posterior odds for multivariate normal classification
J.R.S.S. B 26 pp69-76

Gelsema & Eden (1980)
Mapping algorithms in Ispahan
Pattern recognition 12 pp127-136

Gilbert E.S. (1968)
On discrimination using qualititive variables
J. Amer. Stat. Assoc. 63 p 1399

Gilbert E.S. (1969)
The effect of unequal variance-covariance matrices on Fisher's discriminant function
Biometrics 25 pp 505-516

Gimlin D.R. & Ferrell D.R. (1974)
A K-K' error correcting procedure for non-paramtetric imperfectl supervised learning
IEEE trans Syst Man & Cyber pp304-306

Goldstein M. & Dillon W.R>. (1978)
Discrete discriminant analysis
Wiley
New York

Gordon J.W. (1877)
On certain molar movements of the human body produced by the circulation of the blood
J. Anat. 11 pp533-536

Goshi S.K. & Shyamasunder R.K. (1983
A linear time algorithm for obtaining the convex hull of a simple polytope
Pattern recognition 16 pp587-592

Gower J.C. (1966)
Some distance properties of latent root & vector methods used in multivariate analysis
Biometrika 53 pp325-338

Gower J.C. (1968)
Adding a point to vector diagrams in multivariate analysis
Biometrika 55 pp582-585

Graham R.L. (1972)
An efficient algorithm for determining the convex hull of a finite planar set
Inform Proc Lett pp132-133

Grebenarov (1973)
in
Ballistocardiographic methods and cardiovascular dynamics
Procedings of the Congress in W. Sofia
Ed. A. Talakov
Karger

Green P. rman B.W. (1979)
Constructing the convex hull of a set of points in the plane
Computer Journal 22 pp262-266

Greville (1968)
Data fitting by spline functions
MRC Tech Summ Rep 893
Math Res Centre, US Army
University of Wisconsin

Guseman L.F., Peter H.F.s B.C. & Walker (1975)
On minimising the probability of misclassification for linear feature selection
Annals of Statistics 3 pp661-668

Guttman I. (1970)
Statistical tolerance regions
Griffin
London

Haff L.R. (1980)
Empirical Bayes estimation of the multivariate normal covariance matrix
Annals Statist. 8 pp 586-597

Hand D.J. (1981)
Branch and bound in statistical data analaysis
The statistician 30 pp 1-13

Hand D.J. (1981)
Discrimination and classification
Wiley New York

Hanka R. (1978)
Computerised classification of Ballistocardiograms in
Biosigma 78
International conference on signals & images in medicine and biology Paris

Haralick R.M., Shanmugan K. & Din I.stein (1973)
textural features for image classification
IEEE trans syst man cyber SMC-3 pp610-621

Harrison W.K. & Talbot S.A. (1967)
in
Ballistocardiology & Cardiac Performance
Ed. Noordergraaf & Pollack
Basel

Hart P.E. (1968)
The condensed nearest neighbour rule
IEEE trans on Information theory IT-14 pp515-516

Hartigan J.A. (1975)
Clustering algorithms
John Wiley & sons New York

Hartley H.O. & Rao J.N.K. (1968)
Classification and estimation in analysis of variance problems
Review of international Statistical Inst. 36 pp141-147

Hayes J.G. & Haliday J. (1974)
The least squares fitting of cubic spline surfaces to general data sets
J. Inst Math Appl 14 pp89-103

Healy M.P. & Parrish E.A. (1979)
An optimal method of ordering the basis vectors in the truncated Karhunen-Loeve expansion using the Fisher linear discriminant.
Unpublished

Hellman M.E. (1970)
The nearest neighbour classification rule with a reject option
IEEE Trans. Syst. Sci. Cybern. SSC-6 pp 179-185

Henderson (1905)
The man-movements of the circulation as shown by a recoil curve
Amer J. Physiol. 14 pp287-298

Hills M. (1967)
Discrimination and allocation with discrete data
J.R.S.S. C 16 pp237-250

Hoaglin D.C. & Andrews D.F. (1975)
The reporting of computer based results in statistics
Amer Stat 29 pp122-126

Hosmer D.W. (1973)
A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample.
Biometrics 29 pp761-779

Hotelling H. (1933)
Analysis of a complex of statistical variables into principal components
J Educational Psychology 24 pp417-441

Jain A.K. & Waller W.G. (1978)
On the optimal number of features in the classification of multivariate Gaussian data.
Pattern Recohnition 10 pp365-374

Jaramlokova-Nicolova (1973)
in
Ballistocardiographic methods & cardiovascular dynamics
Procedings of the congress in Sofia
Ed. A. Talakov
Karger

Jardin N. & Sibson R. (1971)
Numerical Taxonomy
Wiley, New York

Jarvis R.A. (1973)
On the identification of the convex hill of a finite set of points in the plane.
Inform Proc Lett 18-21

Jhingram S.G., Jordan L., Jahns M.F. & Haynie T.P. (1971)
Liver scintigrams compared with alkaline phospatase & BSP determinations in the detection of metastatic carcinoma
J. Nucl. Med. 12 pp227-230

John S. (1970)
On identifying the population of origin of each observation in a mixture of observataions from two normal populations
Technometrics 12 pp 553-563

Kabe D.G. (1963)
Some results on the distribution of two random matrices in classification procedures
Annals of Math. Statistics 34 pp181-185

Kailath T. (1967)
The divergence and Bhattacharyya distance measures in signal detection
IEEE Trans COM-15 p52

Kanal L. (1974)
Patterns in Pattern Recognition 1968-1974
IEEE Trans Inform Theory IT-20 pp697-722

Kashyap R.L. (1970)
Algorithms for pattern classification
in
Adaptive learning and pattern recognition systems
Ed. Mendel J.M. & Fu K.J.
Academic Press, New York

Kazakos D. (1977)
Recursive estimation of prior probabilities using a mixture
IEEE trans inform theory IT-23 pp203-211

Kendall M.G. (1966)
Discrimination and Classification
in
Proc of an international symposium on multivariate analysis
Academic Press, New York

Kendall M.G. & Stuart A. (1966)
The advances theory of statistics
Griffin London

Kennedy W.J. & Gentle J.E. (1980)
Statistical computing
Dekker New York

Kittler J. (1975)
Mathematical methods of feature selection in pattern recognition
Internat. J. Man-Machine Studies 7 pp 609-637

Kittler J. (1978)
Feature selection methods based on the Karhunen-Loeve expansion
in
Pattern recognition: Theory and Application
Ed. Fu K.S. & Winston A.B.
Noordhoff

Kittler J. (1978)
Feature set search algorithms
In Pattern recognition and signal processing
Ed Chen C.H.
Sijthoff and Noordhoff Netherlands

Kittler J. & Young (1973)
A new approach to feature selection based on the Karhunen-Loeve expansion
Pattern recognition 5 pp335-352

Koplowitz J. & Brown T.A. (1979)
The weighted nearest neighbour rule for class dependent sample sizes
IEEE trans inform theory IT-25 PP617-619

Kullback S. (1959)
Information theory and statistics
Wiley New York

Lachenbruch P.A. (1965)
Estimation of error rates in discriminant analysis
Ph D dissertation, University of California, Los Angeles

Lachenbruch P.A. (1974)
Discriminant analysis where the initial samples are misclassified II: Non-random misclassification models
Technometrics 16 pp419-424

Lachenbruch P.A. (1975)
Discriminant Analysis
Hafner Press

Lachenbruch P.A. (1979)
Note on initial misclassification effects on the quadratic discriminant function
Technometrics 21 pp129-132

Lachenbruch P.A. & Goldstein (1979)
Discriminant Analysis
Biometrics 35 pp69-85

Lachenbruch P.A. & Mickey M.R. (1968)
Estimation of error rates in discriminant analysis
Technometrics 10 pp1-11

Lachenbruch P.A., Sneenriger C. & Revo L.T. (1973)
Robustness of the linear & quadratic discriminant function to certain types of non-normality
Commun. Stat. 1 pp 39-57

Lainiotis D.G. & Park S.K. (1972)
Feature extraction criteria: comparison and evaluation
Proc 5th Hawai. Internat. Conf. Syst. Sci.

Ledley R.S. Ed. (1968)
Pattern Recognition
Pergamon, New York

Liddell D. (1977)
Multivariate response in more than one sample
The Statistician 26 pp1-15

Lindeman R.H., Merenda P.F. & Gold R.Z. (1980)
Introduction to bivariate and multivariate analysis
Scott, Foresman & Co.

Lissack T. & Fu K.S. (1973)
Error estimation and its application to feature extraction in pattern recognition
Purdue Univ. Lafayette Ind. Rep. TR-EE73-25

Ludbrook J., Slavotinek A.H. & Ronai P.M. (1972)
Observer error in reporting on liver scans for space occupying lesions
Gastroenterology 62 pp1013-1019

Marks S. & Dunn O.J. (1974)
Discriminant functions when covariance matrices are unequal
J. Amer. Stat. Assoc. 69 pp 555-559

Marriott F.H.C. (1975)
Separating mixtures of normal distributions
Biometrics 31 pp767-769

McKay (1976)
Simultaneous procedures in discriminant analysis involving two groups
Technometrics 18 pp 47-53

McLachan G.J. (1972)
Asymptotic results for discriminant analysis when initial samples are misclassified
Technometrics 14 pp415-422

McLachan G.J. (1975)
Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis
J.A.S.A. 70 pp365-369

McLachan G.J. (1980)
The classification and mixture maximum likelihood approaches to cluster analysis
in
The handbook of statistics
Ed P. Krishnaiah
North Holland Amsterdam

Mendell J.M. & Fu K.S. (1970)
Adaptive learning and pattern recognition systems: Theory and applications
Academic, New York

Moore D.H. (1973)
Evaluation of five discrimination procedures for binary variables
J. Amer. Stat. Assoc. 68 pp 399-404

Moran M.A. & Murphy B.J. (1979)
A clear look at two alternative methods of statistical discrimination
Applied Statistics 28 pp223-232

Mucciardi A.N. & Gose E.E. (1971)
A comparison of 7 techniques for choosing subsets of pattern recognition properties
IEEE Trans Computers C-20 pp1023-1031

Murray & Titterington (1978)
Estimation problems with data from a mixture
Applied Statistics

Nelder J.A. & Mead R. (1965)
A simplex method for function minimisation
The Computer Journal 7 pp308-313

Okamoto M. (1963)
An asymptotic expansion for the distribution of the linear discriminant function
Annals of Math Statistics 34 pp1286-1301

Okamoto M. (1968)
Correction to: an asymptotic expansion for the distribution of the linear discriminant function
Annals of Math Statistics 39 pp1358-1380

Olsen D.R. & Fukunaga K. (1973)
Representation of nonlinear data surfaces
IEEE Trans on Computers C-22 pp912-922

O'Neill T. (1978)
Normal discrimination with unclassified observations
J.A.S.A. 73 pp821-826

Orwlawski M. (1983)
On the conditions for the success of Sklansky's hull algorithm
Pattern recognition 16 pp579-586

Parzen E. (1962)
On the estimation of a probability density function and mode
Annals Math Statistics 33 pp1065-1076

Patrick E.A. (1972)
Fundementals of pattern recognition
Prentice Hall, New Jersey

Patrick E.A. & Fisher F.P. (1969)
Non-parametric feature selection
IEEE Trans IT-15 p 577

Peck R. & Van Ness J. (1982)
The use of shrinkage estimators in linear discriminant analysis
IEEE Trans. Pattern Anal. & Mach. Intell. PAM-4 pp 530-537

Peterson D.W. & Mattsion R.L. (1966)
A method for finding linear discriminant functions for a class of performance criteria
IEEE Trans Inform Theory pp380-387

Preparta F.P. & Hong S.J. (1977)
Convex hulls of finite sets of points in 2 & 3 dimensions
Comm of the ACM 20 pp87-93

Queensbury C.P. & Gessaman M.P. (1968)
Nonparametric discrimination using tolerance regions
Annals Math. Statist. 39 pp 664-673

Rao C.R. (1944)
Tests with discriminant functions in multivariate analysis
Sankhya 7 pp407-413

Rao C.R. (1970)
Inference on discriminant function coefficients
In Essays in probability and statistics
Ed Rose R.C. et al
University of North Carolina Press Chapel Hill pp 557-602

Roberts S. (1984)
AS199 A branch & bound algorithm for determining the optimal feature subset of size S
Applied Statistics 33 pp236-241

Rosenfeld A. (1984)
Image analysis: Problems, progress & prospects
Pattern recognition 17 pp3-12

Sammon J.W. (1968)
On line pattern analysis and recognition systems (OLPARS)
Rome Air Develop Center
Tech-Rep TR-68-263

Sammon J.W. (1969)
A nonlinear mapping for data structure analysis
IEEE Trans Computers C-18 pp401-409

Scherer U., Bull U., Roth R., Eisenberg J., Schildberg F.W., Meister P. & Lissner J. (1978)
Computerised tomography & nuclear imaging of the liver
European J. of Nucl. Med. 3 pp71-80

Scott A.J. & Symmons M.J. (1971)
Clustering methods based on likelihood ratio criteria
Biometrics 27 pp387-397

Shanmugan K. & Breipol A.M. (1971)
An error correcting procedure for learning with an imperfect teacher
IEEE Trans Syst. Man & Cybern. pp223-229

Silvey S.D. (1970)
Statistical Inference
Penguin, London

Sklansky J. (1972)
Measuring concavity on a rectangular mosaic
IEEE Trans Computers C-21 pp1355-1364

Torgensen W.S. (1958)
Theory and methods of scaling
Wiley, New York

Tou J.T. & Heydorn R.P. (1967)
Some applications to optimum feature selection
in
Computers & Information Sciences vol II
Ed. Tou J.T.
Academic Press New York

Tou J.T. & Gonzalez R.C.
Pattern Recognition
Addison-Wesley Mass.

Toussaint G.T. (1978)
The convex hull as a tool in pattern recognition
in
Proc AFOSR workshop in communications theory & applications
Provicetown Mass

Tukey J.W. (1947)
Non-parametric estimation II statistically equivalent blocks
- the continuous case
Ann Math Stat 18 pp529-539

Ullman J.R. (1973)
Pattern recognition techniques
Crane, Russak & Co, New York

Vadja I. (1970)
Note on discrimination information and variation
IEEE Trans IT-16 p 771

Van Ness J.W. (1979)
On the effects of dimension in discriminant analysis for
unequal covariance populations
Technometrics 21 pp 119-127

Van Ness J. (1980)
On the dominance of non-parametric Bayes Rule discriminant
algoirthms in higher dimensions
Pattern recognition 12 pp355-368

Van Ness J.W. & Simpson C. (1976)
On the effects of dimension in discriminant analysis
Technometrics 18 pp 175-187

Van Ryzin J. Ed. (1977)
Classification and clustering
Academic Press New York

Verhagen C.J.D.M. (1975)
Some general remarks about pattern recognition; Its definition
Its relation with other discipines: A literature survey
Pattern Recognition 7

Vilmansen T.R. (1973)
Feature extraction with measures of probabilistic dependence
IEEE Trans C-22 p 381

Vido I., Hundeshazen H., Becker H. & Schmidt F.W. (1975)
Verleich laparoskopischer und szintigraphischer befunde
bei chronischer Hepatitis, Leberzirrhose und lebertumsren
Dtch Med Wochenschr 100 pp129-132

Wald A. (1944)
On a statistical problem arising in the classification of
an individual into one of two groups
Ann Math Stat 15 pp145-162

Wald A. & Wolfowitz J. (1946)
Tolerance limits for a normal distribution
Ann Math Stat 17 pp208-215

Watanabe S. (1965)
Karhunen-Loeve expansion and factor analysis - theoretic
remarks and applications
Proc. 4th Prague conf. Information theory

Watanabe S. (1972)
Frontiers of pattern recognition
Academic press, New York

Wegman E.J. (1972)
Non-parametric density estimation
Technometrics 14 pp 533-546

Weissberg A. & Beatty G.H. (1960)
tables of tolerance-limit factors for normal distributions
Technometrics 2 pp483-500

Weiner J. & Dunn O.J. (1966)
Elimination of variates in linear discimination problems
Biometrics 22 p268

Welch B.L. (1939)
Note on disrciminant functions
Biometrika 31 pp 218-220

Wezka J.S., Dyer C.A. & Rosenfeld A. (1976)
A comparitive study of texture measures for terrain classification
IEEE Trans Syst Man Cyber SMC-6 pp269-285

Wilks S. (1941)
Determination of sample sizes for setting tolerance limits
Ann Math Stat 12 pp91-96

Wilks S. (1962)
Mathematical statistics
Wiley New York

Wilson D.L. (1972)
Asymptotic properties of nearest neighbour rules using edited data
IEEE trans Syst Man Cyber SMC-2 pp408-420

Wold S. (1976)
Pattern recognition by means of disjoint principal ccomponent models
Pattern recognition 8 p 127

Wolfe J.H. (1970)
Pattern clustering by multivariate mixture analysis
Mult. Behav. Res. 5 pp329-350

Zubin J. (1938)
A technique for measuring likemindedness
J. Abnorm Soc Psychol 33 pp508-516